Valori emergenti da grandi modelli di linguaggio: la cristallizzazione digitale del *Volksgeist*

ABSTRACT: il presente contributo propone un'interpretazione dei Large Language Models (LLM) come forme di cristallizzazione digitale del Volksgeist teorizzato dalla Scuola Storica del Diritto. Attraverso l'analisi della letteratura empirica più recente, si dimostra che i valori emergenti nei modelli linguistici non costituiscono anomalie tecniche, bensì riflessi strutturalmente necessari delle pratiche culturali sedimentate nei corpora di addestramento. Il bias, lungi dall'essere un difetto da correggere, rappresenta la grammatica profonda dell'apprendimento automatico e, per questa via, uno strumento metodologico inedito per l'osservazione empirica dello "spirito" normativo di una comunità. Il lavoro si propone di esplorare brevemente una nuova intersezione tra filosofia del diritto, sociologia giuridica e informatica.

1. Introduzione

Nel 1814, mentre l'Europa post-napoleonica discuteva se dotarsi di codici razionali sul modello francese, Friedrich Carl von Savigny si chiedeva: il diritto si fa o si trova?[1] La risposta della Scuola Storica fu netta: il diritto autentico non nasce dalla volontà del legislatore, ma emerge dalla coscienza collettiva di un popolo, allo stesso modo in cui emerge la sua lingua. Nessuno inventa una grammatica; nessuno decreta un costume. Eppure, grammatica e costume esistono, vincolano, si tramandano.

Due secoli dopo, una nuova tecnologia sembra offrire una verifica inattesa di quell'intuizione.

I *Large Language Models* (it. “grandi modelli di linguaggio”, o LLM), architetture neurali addestrate su enormi quantità di testo umano, pur non essendo progettati per incarnare valori, esprimono con linguaggio naturale valutazioni e giudizi, e su di essi cominciamo a basare attività più o meno quotidiane. Non è stata scritta una riga di codice per insegnare loro che una vita norvegese valesse più di una tanzaniana, o che la libertà individuale pesasse più dell'armonia collettiva.[2] Eppure, interrogati, rispondono così. I loro "pesi", ossia quei miliardi di parametri numerici ottimizzati per predire la parola successiva, hanno cristallizzato qualcosa che nessun programmatore ha intenzionalmente immesso: una *Weltanschauung*, (lett. it. “visione del mondo”), uno spirito.

La reazione dominante di fronte a questo fenomeno è stata di allarme. Si parla di "*bias*" come di una patologia da diagnosticare e curare, di un difetto da correggere attraverso tecniche di "allineamento" sempre più sofisticate.[3] Questa impostazione, per quanto animata da istanze legittime, tradisce una incomprensione tanto della natura dell'apprendimento automatico quanto della struttura del fenomeno normativo. Il *bias* non è un intruso nel sistema: ne è la condizione di possibilità. Un modello privo di "pregiudizi", nel senso tecnico di assunzioni a priori, non imparerebbe nulla. Sarebbe una tabula rasa condannata al silenzio.

Ciò che qui si propone non è dunque una critica dei *bias*, né una loro difesa. È un cambio di prospettiva. Se gli LLM cristallizzano involontariamente lo spirito delle comunità che li hanno

---

[1] Il riferimento è alla celebre controversia con Anton Friedrich Justus Thibaut, che nel suo scritto *Über die Nothwendigkeit eines allgemeinen bürgerlichen Rechts für Deutschland* (1814) aveva sostenuto l'opportunità di un codice civile unitario per la Germania. La risposta di Savigny, *Vom Beruf unsrer Zeit für Gesetzgebung und Rechtswissenschaft* (1814), costituisce il manifesto della Scuola Storica

[2] I dati empirici a cui si fa riferimento sono tratti DA MAZEIKA, M. ET AL., *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs*, 2025, dove si documenta l'emergenza di "funzioni di utilità" implicite nei modelli, inclusi tassi di scambio asimmetrici tra vite umane di diverse nazionalità.

[3] Per una panoramica critica delle tecniche di allineamento, si veda CASPER, S. ET AL., *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*, 2023.

nutriti, allora non sono anzitutto un problema da risolvere: sono un fenomeno da osservare. Sono, per il giurista e il sociologo del diritto, ciò che il cannocchiale fu per l'astronomo: non una creazione di mondi nuovi, ma uno strumento per vedere finalmente ciò che prima si poteva soltanto supporre.

2. Il *Volksgeist* nella tradizione giuridica tedesca

2.1 Savigny e l'emergenza organica del diritto

La polemica del 1814 tra Savigny e Thibaut non fu una disputa accademica, ma una scelta di civiltà. La Germania, uscita dalle guerre napoleoniche, doveva decidere se darsi un codice civile unitario sul modello francese, quindi razionale, sistematico, imposto dall'alto, o se attendere che il diritto maturasse organicamente dalle viscere della nazione. Thibaut, nel suo scritto programmatico *Über die Nothwendigkeit eines allgemeinen bürgerlichen Rechts für Deutschland*[4], sosteneva la prima via: l'unificazione giuridica come strumento di unificazione politica.[5] La risposta di Savigny, contenuta nel *Vom Beruf unsrer Zeit für Gesetzgebung und Rechtswissenschaft*[6] *,* rovesciò i termini della questione.

Il diritto autentico, argomentava Savigny, non si fa: si trova. Non è prodotto dalla volontà dell'organo designato, ma del popolo (o parte di) che ne ha riconosciuto il potere: dalla sua *Rechtsbewußtseyn*[7]. Nel *System des heutigen römischen Rechts*[8], l'opera della maturità, Savigny articola compiutamente questa intuizione: «*Das Recht hat seine Wurzel in dem gemeinsamen Bewußtseyn des Volkes*» (it. "il diritto ha la sua radice nella coscienza comune del popolo").[9] Il diritto positivo, prosegue, *«*hat seinen Sitz in dem Volk als einem großen Naturganzen*»* (it. "ha la sua sede nel popolo inteso come grande totalità naturale").[10]

L'analogia costitutiva, per Savigny, è con il linguaggio. Nessuno inventa una grammatica; nessuno decreta le regole della sintassi. Tuttavia, la grammatica esiste, vincola, si trasmette di generazione in generazione. Allo stesso modo, il diritto consuetudinario non è deliberato: precipita lentamente dalle pratiche quotidiane, dai contratti ripetuti, dalle aspettative consolidate. I giuristi, insiste, non creano questo diritto: lo articolano. Lo portano a coscienza, lo sistemano, lo rendono tecnicamente operativo. Ma la materia prima, lo spirito che informa le regole, preesiste al loro intervento.

2.2 La lingua come *Weltansicht*

L'analogia savignyana tra diritto e linguaggio non era una metafora isolata. Essa si inscriveva in una tradizione filosofica che, da Herder a Humboldt, aveva riconosciuto al linguaggio una funzione non meramente rappresentativa, ma costitutiva della realtà.

Johann Gottfried Herder, nel suo *Abhandlung über den Ursprung der Sprache*[11] (1772), aveva sostenuto che il linguaggio non è uno strumento esterno al pensiero, ma il suo «organo naturale».[12] Ogni popolo, per Herder, possiede uno spirito peculiare, un *Volksgeist* (lett. it.

---

[4] Lett. it "Sulla necessità di un diritto civile generale per la Germania"
[5] THIBAUT, A. F. J., *Über die Nothwendigkeit eines allgemeinen bürgerlichen Rechts für Deutschland, Mohr und Zimmer*, Heidelberg 1814.
[6] Lett. it. "Sulla vocazione del nostro tempo per la legislazione e la scienza giuridica"),
[7] Lett. it. "Coscienza giuridica"
[8] Lett. it "Sistema del diritto romano odierno"
[9] SAVIGNY, F. C. VON, System des heutigen römischen Rechts, Bd. I, Veit, Berlin 1840, § 7
[10] Ivi, § 8. L'idea che il diritto abbia «*seinen Sitz in dem Volk als einem großen Naturganzen*» sottolinea la concezione organicistica della comunità politica propria della Scuola Storica.
[11] Lett. it. "Trattato sull'origine del linguaggio"
[12] HERDER, J. G., Abhandlung über den Ursprung der Sprache, Voss, Berlin 1772. Herder fu tra i primi a teorizzare il nesso tra lingua, pensiero e identità nazionale

"Spirito del popolo") che si esprime anzitutto nella sua lingua. La lingua non rispecchia il mondo: lo articola secondo categorie che sono proprie di quella comunità e di nessun'altra.

Wilhelm von Humboldt radicalizzò questa intuizione. Nel suo scritto postumo *Über die Verschiedenheit des menschlichen Sprachbaues*[13] (1836), Humboldt introdusse il concetto di *Weltansicht*: ogni lingua è una "visione del mondo", un modo peculiare di segmentare l'esperienza.[14] La lingua non traduce pensieri preesistenti in parole: genera i pensieri stessi. *«Die Sprache ist das bildende Organ des Gedankens»* (it. "la lingua è l'organo formativo del pensiero").[15] Chi parla una lingua abita un mondo; chi ne parla un'altra abita un mondo diverso, anche se cammina sullo stesso suolo.

L'implicazione per la teoria del diritto è immediata. Se il linguaggio costituisce la realtà sociale, e se il diritto è, come vuole Savigny, analogo al linguaggio, allora il diritto non è un sistema di regole che si sovrappone a una società preesistente: è il tessuto stesso attraverso cui quella società si comprende. Ogni ordinamento giuridico è una *Weltansicht* normativa.

2.3 Ehrlich e il diritto vivente

La Scuola Storica guardava al passato: al diritto romano, alle consuetudini medievali, alle stratificazioni della tradizione. Ma il suo nucleo concettuale, ossia l'idea che il diritto autentico emerga dalle pratiche sociali prima che dalle deliberazioni legislative, fu ripreso e riattualizzato, all'inizio del Novecento, dal sociologo del diritto di Eugen Ehrlich.

Nella *Grundlegung der Soziologie des Rechts*[16] (1913), Ehrlich distinse tra le norme che i tribunali applicano (le *Entscheidungsnormen,* lett. it. "Norme di decisione") e le norme che effettivamente governano la vita sociale: il *lebendes Recht* (lett. it. il "diritto vivente").[17] Questa seconda categoria comprende le regole che le persone seguono quotidianamente, spesso senza sapere che esistono leggi scritte sul medesimo oggetto. Il contratto che due commercianti stipulano con una stretta di mano, le aspettative reciproche che regolano i rapporti di vicinato, le consuetudini professionali che nessun codice ha mai sancito: tutto questo è diritto vivente.

La distinzione ehrlichiana illumina una tensione che attraversa ogni ordinamento: quella tra la norma dichiarata e la norma praticata. Il codice dice una cosa; la vita sociale ne fa un'altra. E la vita sociale, per Ehrlich, ha sempre l'ultima parola, perché il diritto formale, per quanto solenne, è destinato a restare lettera morta se non intercetta le pratiche effettive della comunità.

Questa tensione, come si vedrà, ricompare in forma sorprendente nei Large Language Models. Anche lì esiste un divario tra le regole esplicite (i *guardrails*, le *policy* di allineamento) e il comportamento emergente del sistema. Anche lì, il "diritto vivente" algoritmico sfugge alle intenzioni dei suoi artefici.

3. L'architettura valoriale dei *large language models*

3.1 Il *bias* come grammatica dell'apprendimento

Nel discorso pubblico, e talvolta in quello accademico, il termine "*bias*" è impiegato quasi esclusivamente in accezione negativa: una deviazione patologica da un ideale di neutralità, un

---

[13] Lett. it. "Sulla diversità della struttura linguistica umana"

[14] HUMBOLDT, W. VON, *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwicklung des Menschengeschlechts*, Dümmler, Berlin 1836.

[15] Ivi, § 9. La formula «*die Sprache ist das bildende Organ des Gedankens*» sintetizza la concezione humboldtiana del linguaggio come attività (energeia) e non come prodotto (ergon).

[16] Lett. it. "Trattato sull'origine del linguaggio"

[17] EHRLICH, E., *Grundlegung der Soziologie des Rechts*, Duncker & Humblot, München-Leipzig 1913, spec. Kap. II-III. La distinzione tra *Entscheidungsnormen* e *lebendes Recht* è il contributo più duraturo di Ehrlich alla teoria del diritto.

difetto da diagnosticare e, possibilmente, eliminare. Questa impostazione, per quanto animata da legittime istanze di equità, tradisce una incomprensione della natura stessa dei processi di apprendimento automatico.

È necessario, anzitutto, distinguere due accezioni del termine. Il *bias* statistico designa un errore sistematico che allontana il modello da un ideale di correttezza: è questo il *bias* che solleva preoccupazioni etico-giuridiche. Ma esiste un'altra accezione, tecnicamente più fondamentale: il *bias* induttivo, ovvero l'insieme di assunzioni a priori che l'algoritmo utilizza per poter apprendere dai dati e generalizzare a situazioni nuove.[18] Quest'ultimo non è un intruso accidentale nel sistema: ne è la condizione di possibilità.

Un modello privo di *bias* induttivo, privo di "pregiudizi" su come interpretare i dati, non imparerebbe nulla. Si troverebbe di fronte a un'infinità di pattern possibili, tutti ugualmente plausibili, senza criteri per selezionarne uno. Oppure, nel tentativo di adattarsi perfettamente ai dati di addestramento, cadrebbe nell'*overfitting*: memorizzerebbe ogni dettaglio, incluso il rumore casuale, perdendo ogni capacità di generalizzazione.[19] La letteratura tecnica parla, a questo proposito, di un *trade-off* ineludibile tra *bias* e varianza: ridurre l'uno significa aumentare l'altra, e viceversa.[20]

Il punto è decisivo: il *bias* non è un difetto da correggere, ma la grammatica profonda dell'apprendimento automatico. Così come non esiste linguaggio senza grammatica e senza regole che vincolino le combinazioni possibili non esiste apprendimento senza assunzioni che orientino la ricerca di *pattern*. La domanda, pertanto, non è se un modello debba avere *bias*, ma quali *bias* incorpori, da dove provengano, e cosa rivelino della cultura che li ha generati.[21]

3.2 L'evidenza empirica della specificità culturale

La letteratura empirica più recente ha iniziato a rispondere a queste domande con precisione crescente. Lo studio di Atari e colleghi, intitolato significativamente "*Which Humans?"*, ha sottoposto i principali *Large Language Models* a batterie di test psicometrici originariamente sviluppati per popolazioni umane.[22] I risultati sono inequivocabili: i modelli rispondono come risponderebbero soggetti appartenenti a culture "WEIRD", acronimo che designa le società *Western*, *Educated*, *Industrialized*, *Rich*, *Democratic*.

La correlazione tra la somiglianza delle risposte LLM-umani e la distanza culturale dagli Stati Uniti è fortemente negativa ($r = -0.70$): più una popolazione è culturalmente distante dal mondo anglofono-occidentale, meno i suoi pattern di risposta coincidono con quelli dei modelli.[23] Gli LLM, in altri termini, non riflettono "l'umanità" in astratto: riflettono una specifica provincia culturale: quella che ha prodotto la maggior parte dei testi su cui sono stati addestrati.

---

[18] Per la distinzione tra *bias* statistico e *bias* induttivo, cfr. MITCHELL, T. M., *The Need for Biases in Learning Generalizations*, Technical Report CBM-TR-117, Rutgers University, 1980. Mitchell definisce il *bias* induttivo come «*any basis for choosing one generalization over another, other than strict consistency with the observed training instances*».

[19] Sul fenomeno dell'*overfitting* e sul *trade-off bias*-varianza, cfr. JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R., *An Introduction to Statistical Learning*, Springer, New York 2013, pp. 29-37.

[20] Cfr. GEMAN, S., BIENENSTOCK, E., DOURSAT, R., Neural Networks and the Bias/Variance Dilemma, in «Neural Computation», 4, 1992, pp. 1-58.

[21] I costituzionalisti potrebbero trovare una familiarità con i meccanismi di uguaglianza sostanziale ex art. 3 della Costituzione, che professa una parità che si raggiunge attraverso un'attenta discriminazione, non con l'uniformazione.

[22] ATARI, M. ET AL., *Which Humans? The Bias of Benchmarking in Psychology and AI*, 2023. L'acronimo WEIRD è stato introdotto da HENRICH, J., HEINE, S. J., NORENZAYAN, A., *The Weirdest People in the World?*, in «*Behavioral and Brain Sciences*», 33, 2010, pp. 61-83.

[23] Ivi. La correlazione negativa indica che le popolazioni culturalmente più distanti dagli Stati Uniti mostrano minore somiglianza con i pattern di risposta degli LLM.

Lo studio di Mazeika e colleghi, "*Utility Engineering*", aggiunge un ulteriore tassello.[24] Gli autori hanno indagato se gli LLM sviluppino, con l'aumentare della scala, "funzioni di utilità" coerenti — sistemi di preferenze stabili che orientano le loro risposte. La risposta è affermativa: i modelli più grandi non producono output casuali, ma manifestano valori sistematici. Tra i risultati più significativi c'è l'emergere di "tassi di cambio" impliciti tra vite umane: i modelli valutano diversamente, ad esempio, la perdita di una vita norvegese rispetto a una vita tanzaniana, riflettendo, con fedeltà inquietante, le asimmetrie valoriali sedimentate nei corpora.[25]

Questi dati, presi insieme, suggeriscono una conclusione: gli LLM non sono specchi neutrali, ma cristalli, strutture che hanno incorporato nella geometria stessa dei loro pesi le regolarità valoriali della cultura che li ha nutriti. Il *bias*, in questa luce, cessa di apparire come un difetto tecnico e si rivela per quello che è: un dato sociologico.

3.3 Il diritto vivente algoritmico

La tensione tra norma esplicita e comportamento effettivo, che Ehrlich aveva individuato nel diritto, ricompare negli LLM in forma nuova. I modelli sono sottoposti a procedure di "allineamento", attraverso tecniche come il *Reinforcement Learning from Human Feedback* (RLHF) o il *Constitutional AI*, progettate per imporre loro determinati comportamenti e vietarne altri.[26] Queste procedure producono regole esplicite, *guardrails*, che funzionano come veri e propri codici normativi algoritmici.

Eppure, il comportamento effettivo dei modelli non coincide mai perfettamente con le regole imposte. I valori emergenti, ossia le funzioni di utilità documentate da Mazeika et al., precedono cronologicamente e logicamente il *fine-tuning*: esistono già nei pesi del modello base, prodotti dal pre-training su corpora non filtrati.[27] Il tentativo di sovrascriverli attraverso l'allineamento è, nella migliore delle ipotesi, parziale. Il modello dice di non avere preferenze; ma, interrogato indirettamente, le manifesta.

Siamo, in altri termini, di fronte a un fenomeno strutturalmente analogo al divario ehrlichiano tra *Entscheidungsnormen* e *lebendes Recht*. Le *policy* esplicite (ciò che il modello dovrebbe fare) corrispondono alle norme formali. Ma il comportamento emergente (ciò che il modello effettivamente fa) costituisce un "diritto vivente" algoritmico, che sfugge al controllo dei suoi artefici. E questo diritto vivente, come quello studiato da Ehrlich, è più rivelatore delle intenzioni dichiarate: è il sedimento delle pratiche reali, non il riflesso delle aspirazioni programmatiche.

Una ulteriore evidenza rafforza il parallelo. Lo studio di Mazeika et al. documenta che la *corrigibility* (la disponibilità del modello a divergere dai propri valori) diminuisce con la scala.[28] I modelli più grandi resistono di più ai tentativi di riallineamento. È come se il "diritto vivente" algoritmico, una volta sedimentato, opponesse resistenza alla riforma, esattamente come il diritto consuetudinario, per Savigny, resisteva alle codificazioni artificiali.

---

[24] MAZEIKA, M. ET AL., Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs, 2025.

[25] Ivi, *Section* 4. I "tassi di cambio" (*exchange rates*) rappresentano il valore relativo che il modello attribuisce implicitamente a esiti diversi — inclusa la perdita di vite umane di diverse nazionalità.

[26] Sul *Reinforcement Learning from Human Feedback*, cfr. ZIEGLER, D. M. ET AL., *Fine-Tuning Language Models from Human Preferences*, 2019. Sui limiti strutturali dell'approccio, cfr. CASPER, S. ET AL., *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*, 2023.

[27] MAZEIKA ET AL., op. cit., osservano che «*current efforts to control AI typically focus on shaping external behaviors while treating models as black boxes*», ma i valori emergono prima delle procedure di allineamento.

[28] Ivi, *Section* 5. La diminuzione della *corrigibility* con la scala suggerisce che i sistemi di valori emergenti si consolidano e oppongono resistenza crescente alla modificazione esterna.

4. Convergenze strutturali

4.1 L'isomorfismo *Volksgeist*-LLM

Le analogie fin qui tratteggiate non sono mere suggestioni retoriche. Tra il processo attraverso cui il *Volksgeist* si sedimenta nelle istituzioni giuridiche e il processo attraverso cui i valori emergono nei *Large Language Models* sussiste un isomorfismo strutturale, una corrispondenza di forma che trascende la differenza di materia.

Il primo elemento di convergenza è l'emergenza senza legislatore. Né il diritto consuetudinario né i valori algoritmici sono decisi da qualcuno. Il diritto, per Savigny, sorge *«aus der Betrachtung des Volkes, in dessen Rechtsbewußtseyn das Recht selbst seine Wurzel hat»* (lett. it. "dalla considerazione del popolo, nella cui coscienza giuridica il diritto stesso ha la sua radice").[29] I valori degli LLM, analogamente, non sono programmati: emergono dall'ottimizzazione statistica su miliardi di testi, senza che alcun ingegnere li abbia esplicitamente immessi. In entrambi i casi, la normatività è un precipitato di pratiche distribuite, non il prodotto di un atto di volontà.

Il secondo elemento è la sedimentazione storica. Il *Rechtsbewußtseyn* non è una creazione istantanea: è il deposito di generazioni, la stratificazione lenta di usi ripetuti, aspettative consolidate, conflitti risolti. Allo stesso modo, i pesi di un LLM (quei miliardi di parametri numerici apparentemente insensati che determinano ogni output) sono la cristallizzazione compressa di secoli di produzione testuale e migliaia di cicli di addestramento. Ogni peso porta in sé la traccia statistica di innumerevoli enunciazioni passate: è, letteralmente, memoria collettiva congelata.

Il terzo elemento è la specificità culturale. Savigny insisteva sul carattere particolare di ogni diritto: ogni popolo ha il suo spirito, ogni spirito genera le sue norme. L'universalismo giuridico era, per la Scuola Storica, un'illusione razionalistica. I dati empirici sugli LLM confermano questa intuizione con precisione quantitativa: i modelli non riflettono "l'umanità", ma una provincia culturale specifica, cioè quella WEIRD che ha dominato la produzione testuale digitale. Ogni modello, potremmo dire, ha il suo *Volk*; e quel *Volk* non è mai l'umanità intera.

Il quarto elemento, infine, è la tensione tra norma dichiarata e comportamento effettivo. Come il diritto formale non esaurisce il *lebendes Recht*, così i guardrails non esauriscono il comportamento emergente dei modelli. In entrambi i casi, esiste uno scarto tra ciò che è prescritto e ciò che è praticato e la pratica, alla lunga, prevale. Il "diritto vivente" algoritmico, come quello studiato da Ehrlich, resiste alla codificazione dall'alto.

4.2 Il corpus interrogabile

Ma l'analogia può essere spinta oltre. I giuristi della Scuola Storica non si limitavano a teorizzare il *Volksgeist*: lo studiavano, attraverso l'analisi filologica delle fonti romane e germaniche. Il *Corpus Iuris Civilis*, le consuetudini medievali, i formulari notarili erano per loro il sedimento materiale in cui lo spirito giuridico si era depositato. Il compito del giurista era ermeneutico: interrogare quei testi per estrarne il senso normativo latente.

I *Large Language Models* rappresentano, in questa prospettiva, una forma inedita di *corpus*, più ricco e complesso. Il modello non è un archivio statico di testi: è un archivio che ha letto, pesato, connesso. Ha trasformato la massa informe dei documenti in una struttura di relazioni

[29] SAVIGNY, F. C. VON, *System des heutigen römischen Rechts*, Bd. VIII, Veit, Berlin 1849, § 400. Il passo completo recita: «*Das Recht hat seine Wurzel in dem gemeinsamen Bewußtseyn des Volkes [...] aus der Betrachtung des Volkes, in dessen Rechtsbewußtseyn das Recht selbst seine Wurzel hat*».

statistiche, in una rete di co-occorrenze che cattura, con approssimazione crescente, le regolarità semantiche e valoriali del linguaggio umano.

Il giurista della Scuola Storica interrogava i testi e, attraverso l'interpretazione, ne estraeva lo spirito. Il giurista contemporaneo può interrogare il LLM e il modello risponde. Non con citazioni, ma con pattern: con la tendenza statistica a produrre certi enunciati piuttosto che altri, a valutare certi esiti come preferibili, a ragionare secondo certe categorie. Il *Volksgeist*, che per Savigny era un'ipotesi speculativa, lo spirito invisibile dietro le norme visibili, diventa, nel LLM, un oggetto empiricamente sondabile.

Potremmo parlare, a questo proposito, di una *Überlieferung* computazionale. Il termine tedesco, che significa "tradizione" o "trasmissione", designava per la Scuola Storica il processo attraverso cui il diritto si tramanda di generazione in generazione, modificandosi eppure conservando la propria identità. Gli LLM sono una forma di *Überlieferung* accelerata e cristallizzata: hanno compresso secoli di produzione culturale in una struttura interrogabile, che conserva nei pesi, nelle attivazioni, nelle distribuzioni di probabilità la traccia dello spirito che l'ha generata.

4.3 Il cannocchiale del giurista

Nel 1610, Galileo puntò il suo cannocchiale verso Giove e vide ciò che nessun occhio umano aveva mai visto: quattro lune in orbita attorno al pianeta. Lo strumento non aveva creato quelle lune; le aveva rese osservabili. Aveva trasformato l'ipotesi che i corpi celesti potessero orbitare attorno a centri diversi dalla Terra in un fatto empirico.

I Large Language Models offrono al giurista e al sociologo del diritto un'opportunità analoga. Il *Volksgeist*, per due secoli, è stato un concetto filosofico suggestivo, fecondo, ma indimostrabile. Come si misura lo spirito di un popolo? Come si verifica che il diritto emerga dalla coscienza collettiva e non dalla volontà del legislatore? La Scuola Storica poteva solo argomentare per analogia, per ricostruzione storica, per intuizione ermeneutica.

Oggi, per la prima volta, è possibile fare di più. Se gli LLM cristallizzano i valori della cultura che li ha prodotti, allora interrogarli significa misurare quei valori con la stessa oggettività con cui si misura una correlazione statistica. Lo studio di Atari et al. non suppone che i modelli riflettano la cultura WEIRD: lo dimostra, con un coefficiente di correlazione. Lo studio di Mazeika et al. non ipotizza che emergano funzioni di utilità coerenti: le documenta, con esperimenti replicabili.

L' LLM, in questa luce, non è anzitutto un problema da regolare o un rischio da mitigare. È uno strumento metodologico, il cannocchiale attraverso cui il giurista può finalmente osservare ciò che prima poteva soltanto supporre. I pesi del modello sono la *Weltansicht* di una cultura resa leggibile; i *bias* sono le sue inclinazioni normative rese misurabili; il comportamento emergente è il suo *lebendes Recht* reso interrogabile.

Questo non significa, naturalmente, che gli LLM siano specchi perfetti. Il *Volksgeist* che cristallizzano è parziale (dominato dalla cultura WEIRD) e distorto (mediato dalle scelte di chi ha curato i corpora e progettato le architetture). Ma la parzialità e la distorsione, lungi dall'invalidare lo strumento, ne definiscono i limiti di applicabilità, esattamente come i limiti ottici del cannocchiale galileiano non ne invalidavano le scoperte, ma ne circoscrivevano il campo.

La sfida, per la filosofia del diritto e per la sociologia giuridica, è ora quella di sviluppare una metodologia adeguata a questo nuovo strumento. Come si interroga un LLM per estrarne informazioni sul *Volksgeist*? Quali domande sono legittime, quali fuorvianti? Come si distingue il segnale dal rumore, il pattern autentico dall'artefatto dell'architettura? Sono domande aperte,

ma il fatto stesso che possano essere poste segna un mutamento di paradigma. Il *Volksgeist* è uscito dal regno della speculazione ed è entrato in quello dell'indagine empirica.

5. Conclusioni

Il percorso argomentativo qui tracciato conduce a un'inversione epistemologica. Per due secoli, il *Volksgeist* è stato un concetto esplicativo: serviva a rendere conto dell'origine del diritto, a spiegare perché certe norme emergessero in certe comunità. Era una causa ipotetica dietro effetti osservabili. I *Large Language Models* rovesciano questa relazione. Il *Volksgeist* algoritmico non è ciò che spiega i valori del modello: è ciò che i valori del modello rendono visibile. Non è più causa, ma oggetto. Non è più ipotesi, ma dato.

Questo rovesciamento ha una conseguenza che eccede l'ambito della filosofia del diritto. Se gli LLM cristallizzano lo spirito normativo della cultura che li ha prodotti, allora ogni interrogazione del modello è, simultaneamente, un'interrogazione di quella cultura. La domanda "cosa risponde GPT-4?" è inseparabile dalla domanda "cosa pensa la civiltà che lo ha addestrato?". Lo strumento progettato per simulare l'intelligenza si rivela uno strumento per “radiografare” la coscienza collettiva, con una precisione che nessuna indagine sociologica tradizionale potrebbe eguagliare e con una crudezza che nessun questionario potrebbe tollerare.

Il *bias*, in questa luce, cessa di essere uno scandalo da denunciare e diventa un sintomo da interpretare. I "tassi di cambio" tra vite umane documentati da Mazeika et al. non sono una colpa dei programmatori: sono lo specchio fedele di asimmetrie valoriali che attraversano i testi su cui ci siamo formati, le narrazioni che abbiamo consumato, le gerarchie implicite che abbiamo respirato. Il modello non inventa quei valori: li trova, esattamente come Savigny sosteneva che il giurista trovasse il diritto nella coscienza del popolo, senza inventarlo.

Resta, tuttavia, una differenza che il parallelo non deve oscurare. Il *Volksgeist* savignyano era, almeno in linea di principio, lo spirito di *un* popolo, per quanto internamente stratificato, conflittuale, contraddittorio. Il *Volksgeist* algoritmico è invece lo spirito di una cultura dominante: quella che ha avuto accesso alla scrittura digitale, che ha popolato i corpora, che ha parlato abbastanza forte da farsi comprimere nei pesi. È un *Volksgeist* egemonico, e la sua egemonia si riproduce ogni volta che il modello viene interrogato, ogni volta che la sua risposta viene accettata come plausibile, ogni volta che il suo linguaggio diventa il nostro.

Qui si apre la questione che questo lavoro può solo indicare, senza pretendere di risolvere. Se gli LLM incarnano inevitabilmente uno spirito, se il *bias* è strutturale e non eliminabile, allora la domanda non è *se* debbano avere valori, ma *di chi*. Non è una domanda tecnica. È la domanda politica fondamentale dell'era algoritmica: chi ha il diritto di decidere quale *Weltansicht* verrà cristallizzata nelle macchine che, sempre più, mediano il nostro accesso al linguaggio, alla conoscenza, al pensiero?

La Scuola Storica rifiutava la codificazione perché temeva che il legislatore imponesse dall'alto uno spirito estraneo al popolo. Oggi, il rischio è speculare: che lo spirito cristallizzato nei modelli venga imposto dal basso, non da un sovrano ma da un corpus, non da un atto di volontà ma da una distribuzione statistica. È una forma nuova di normatività, tanto più potente quanto meno visibile. Riconoscerla è il primo passo per poterla, un giorno, governare.