ChatGPT possiede il *law sense*? Riflessioni a partire da uno studio sperimentale di *argument mining* su decisioni giudiziarie della Corte di Cassazione

Serena Tomasi*

DOES CHATGPT HAVE LAW SENSE? REFLECTIONS ON AN EXPERIMENTAL STUDY OF ARGUMENT MINING IN JUDICIAL DECISIONS OF THE ITALIAN COURT OF CASSATION

ABSTRACT: This article examines the potential and limits of Large Language Models in judicial practice, taking as its test case the reconstruction of legal argumentation within judicial reasoning. Through an argument mining experiment on decisions of the Italian Supreme Court of Cassation, conducted with GPT-4o and interpreted through Philip Bobbitt's theory, the paper asks whether an LLM can grasp not only the textual dimension of a judicial decision, but also its rhetorical form. The results show a concrete usefulness in preliminary tasks of segmentation and textual organization, but also a structural limitation in recognizing ethical argumentation and, more broadly, the law sense required by legal judgment.

KEYWORDS: large language models; argument mining; ethics; law sense; rhetoric

ABSTRACT: Il contributo analizza potenzialità e limiti dei *Large Language Models* nella pratica giudiziale, assumendo come banco di prova la ricostruzione delle argomentazioni giuridiche nella motivazione delle sentenze. Attraverso un esperimento di *argument mining* su decisioni della Corte di Cassazione italiana, condotto con GPT-4o e interpretato alla luce della teoria di Philip Bobbitt, il lavoro verifica se un LLM possa cogliere non solo la dimensione testuale della decisione, ma anche la sua forma retorica. I risultati mostrano un'utilità concreta nelle operazioni preliminari di segmentazione e organizzazione del testo, ma anche un limite strutturale nel riconoscimento dell'argomentazione etica e del *law sense*.

PAROLE CHIAVE: Large Language Models; *argument mining*; etica; *law sense*; retorica

* *Serena Tomasi è ricercatrice a tempo determinato di tipo B in Filosofia del diritto presso l'Università di Trento. Mail: serena.tomasi_1@unitn.it.* *Contributo sottoposto a doppio referaggio anonimo.*

1. Introduzione

Questo scritto intende esaminare le potenzialità e i limiti dei *Large Language Models* nella pratica giudiziale, assumendo come terreno privilegiato di osservazione il problema della ricostruzione delle argomentazioni giuridiche all'interno della motivazione delle sentenze. Il punto non è stabilire se un modello linguistico sia capace di produrre testi giuridicamente verosimili, ma se possa riconoscere e organizzare quelle forme di giustificazione attraverso cui il diritto si presenta come discorso razionale e istituzionalmente vincolato. In altri termini, la questione affrontata è se un LLM e, nel caso specifico, GPT-4o, possa davvero riprodurre la forma retorica del diritto, e non soltanto simularne la regolarità linguistica.

A tale scopo, il contributo prende in esame uno studio sperimentale di *argument mining*[1] applicato a decisioni della Corte di Cassazione, volto a verificare se il modello sia in grado di individuare le unità argomentative, riconoscere i nessi tra passaggi semanticamente affini e classificare le diverse modalità del ragionamento giudiziale, anche quando esse non risultino immediatamente esplicite nel testo[2]. La sfida è particolarmente significativa perché la motivazione giudiziaria non coincide mai integralmente con ciò che appare in forma manifesta: essa contiene argomenti espressi e argomenti impliciti, passaggi di mera ricognizione e passaggi autenticamente giustificativi, elementi descrittivi e snodi valutativi che solo l'interprete è in grado di distinguere alla luce della funzione che essi svolgono nell'economia della decisione[3].

---

[1] L'*argument mining* opera su quelle che la letteratura definisce *argumentative discourse units*, ossia unità minime di analisi argomentativa, e procede poi all'identificazione del loro ruolo e delle relazioni che le collegano, con l'obiettivo di convertire testi non strutturati in strutture argomentative esplicite, ricostruendo non solo ciò che viene sostenuto, ma anche perché viene sostenuto, attraverso l'individuazione di premesse, conclusioni e rapporti di supporto o conflitto. Secondo Lawrence e Reed, sul piano analitico, il campo di ricerca comprende sotto compiti tra loro connessi, quali la segmentazione del testo in unità argomentative, la distinzione tra materiale argomentativo e non argomentativo, la classificazione del ruolo dei segmenti e il riconoscimento delle relazioni argomentative; la sua automazione resta tuttavia difficile, sia per la complessità intrinseca dei testi, sia per la limitata disponibilità di dati annotati di alta qualità. V. J. LAWRENCE, C. REED, *Argument Mining: A Survey*, in *Computational Linguistics*, 45, 4, 2020, 765-818.

[2] La sperimentazione è stata condotta da Serena Tomasi, Jacopo Staiano e Carlotta Giacchetta, afferenti all'Università di Trento, da Raffaella Bernardi, afferente alla Libera Università di Bolzano, e da Barbara Montini, afferente all'Università di Brescia; alla fase di annotazione e validazione hanno inoltre preso parte ricercatori, studenti della Scuola forense di Trento, laureandi della Facoltà di Giurisprudenza dell'Università di Trento e dottorandi in Studi Giuridici Comparati ed Europei dell'Università di Trento. I risultati della ricerca sono stati presentati, in una prima versione, al *12th Argument Mining Workshop* dell'*Association for Computational Linguistics*, svoltosi a Vienna il 31 luglio 2025, e, in una successiva rielaborazione, alla *5th European Conference on Argumentation - Argumentation in the Digital Society* (ECA 2025), tenutasi a Varsavia dal 23 al 26 settembre 2025.

[3] Per una ricostruzione delle teorie dell'argomentazione giuridica di Alexy, MacCormick, Peczenik e Aarnio, v. M. ATIENZA, *Diritto come argomentazione. Concezioni dell'argomentazione*, Napoli, 2019, 5 ss.; sul nesso tra svolta argomentativa, trasformazioni dello Stato costituzionale e ripensamento del ruolo del giudice, v. anche A. ABIGNENTE, Argomentazione giuridica, in U. POMARICI (a cura di), *Atlante di filosofia del diritto*, II, Torino, 2012, 1-36; nonché G. PINO, *Diritti e interpretazione. Il ragionamento giuridico nello Stato costituzionale*, Bologna, 2010. Per un approccio più tecnico-operativo, v. M. ATIENZA, A.L. PRADO, *Cómo analizar una argumentación jurídica*, Quito, 2009. A partire dall'inclusione del discorso giuridico nel discorso pratico generale, per il nesso tra svolta argomentativa, tradizione post-perelmaniana e approdo retorico del ragionamento giuridico, v. S. TOMASI, *L'argomentazione giuridica dopo Perelman. Teorie, tecniche e casi pratici*, Roma, 2020.

Il problema, pertanto, è teorico oltre che tecnico: se il diritto è anche una pratica retorica[4], nel senso che persuade e giustifica mediante forme argomentative riconosciute entro una comunità istituzionale, allora verificare la performance di un LLM nella ricostruzione di tali forme significa interrogare il rapporto stesso tra plausibilità statistica e razionalità giuridica. È su questo crinale che si colloca il presente lavoro: mostrare in quale misura l'intelligenza artificiale possa assistere l'analisi del ragionamento giudiziale e in quale misura, invece, essa riveli un limite strutturale proprio nel confronto con la dimensione implicita, valutativa e normativa dell'argomentazione.

La posta in gioco è duplice. Da un lato, si tratta di valutare se questi sistemi possano offrire un supporto utile nelle operazioni di segmentazione, organizzazione e prima mappatura del testo giudiziario; dall'altro, si tratta di chiarire se tale supporto resti confinato al livello della plausibilità linguistica e della regolarità semantica, oppure se riesca davvero a cogliere ciò che rende un argomento giuridico tale, cioè il suo essere una ragione istituzionalmente situata, retoricamente costruita e orientata. L'assunto di partenza è che il vero banco di prova per i modelli linguistici in ambito giuridico non sia la semplice correttezza formale della risposta, ma la loro capacità (o incapacità) di confrontarsi con la struttura giustificativa delle decisioni: la pratica giudiziale offre, infatti, un osservatorio privilegiato perché rende particolarmente visibile la distanza (e la tensione) tra una produzione linguistica statisticamente plausibile e una motivazione giuridicamente responsabile.

L'analisi si svilupperà in tre passaggi: in primo luogo, verrà definito il quadro teorico di riferimento dell'analisi argomentativa, chiarendo le ragioni per cui è stata scelta la teoria delle modalità argomentative di Philip Bobbitt[5]. Questa opzione non risponde a un'esigenza meramente classificatoria, ma alla necessità di disporre di uno schema capace di descrivere il ragionamento giuridico come articolazione di forme diverse di giustificazione, tra loro distinguibili ma compresenti nella motivazione giudiziale. La teoria di Bobbitt appare particolarmente utile poiché consente di rendere visibili non solo gli argomenti testuali, dottrinali-giurisprudenziali, storici, strutturali e prudenziali, ma anche la specifica rilevanza dell'argomentazione etica.

In secondo luogo, sarà esposto l'esperimento realizzato dal team di ricerca nelle sue diverse fasi metodologiche: la costruzione del corpus, le operazioni di segmentazione del testo, il raggruppamento dei passaggi argomentativi, la loro classificazione secondo la tipologia prescelta e, infine, la validazione dei risultati mediante confronto con annotatori umani. In questa parte saranno presentati anche i risultati della ricerca, sia sotto il profilo quantitativo sia sotto il profilo qualitativo.

Infine, tali risultati saranno ricondotti alla domanda di ricerca da cui il lavoro prende avvio: se e in quale misura un *Large Language Model*, nel caso di specie, GPT-4o, sia capace di ricostruire le argomentazioni giudiziali, tanto esplicite quanto implicite, e dunque di riconoscere non soltanto il profilo enunciativo della decisione, ma anche la sua forma retorica, ossia la struttura giustificativa attraverso cui il giudice organizza e legittima il suo discorso.

[4] Per una lettura del diritto come pratica retorica, v. M. MANZIN, *Argomentazione giuridica e retorica forense. Dieci riletture sul ragionamento processuale,* Torino, 2014; F. PUPPO, *Diritto e retorica*, Torino, 2024; S. TOMASI, *Neil MacCormick e la retorica del diritto*, Torino, 2024.

[5] P. BOBBITT, *Constitutional Fate: Theory of the Constitution*, Oxford, 1982.

2. La teoria di Bobbitt e la grammatica argomentativa del diritto

Il riferimento teorico della ricerca sperimentale è costituito dalla teoria di Philip Bobbitt, che ha rappresentato una proposta originale nel dibattito costituzionalistico contemporaneo, soprattutto nel contesto nordamericano, in cui la riflessione si è sviluppata attorno a due questioni strettamente connesse: la natura dell'interpretazione costituzionale e il fondamento della legittimità del controllo giudiziale di costituzionalità[6]. In tale dibattito si sono intrecciate, spesso senza una netta separazione, due grandi linee di discussione: da un lato, quella relativa al metodo dell'interpretazione costituzionale, e dunque al problema di dove debba essere ricercato il significato della Costituzione e secondo quali criteri esso possa essere correttamente individuato; dall'altro, quella concernente la giustificazione sostanziale dell'ordine costituzionale e, quindi, il tipo di ragioni che devono orientare il giudice quando è chiamato a decidere questioni che investono l'esercizio e i limiti del potere. In una prima prospettiva, il problema viene formulato essenzialmente come problema metodologico: ci si domanda se il significato della Costituzione debba essere rinvenuto anzitutto nel testo in cui essa si esprime, oppure nelle intenzioni originarie di coloro che l'hanno redatta e approvata, oppure ancora nelle strutture di governo che essa istituisce e nei rapporti che organizza tra i poteri pubblici; e, pur nella diversità delle risposte offerte, ciò che accomuna questi orientamenti è l'idea che la questione decisiva consista nell'individuare il criterio interpretativo corretto, vale a dire la via più affidabile per accertare il significato costituzionale[7]. Accanto a questa linea di riflessione, se ne sviluppa però un'altra, la quale tende a relativizzare la centralità del problema metodologico in senso stretto, sul presupposto che la Costituzione non possa essere considerata soltanto come un testo da interpretare o come il deposito di una volontà originaria da ricostruire, ma debba essere compresa anche come il dispositivo attraverso cui una comunità politica cerca di impedire che il potere agisca ingiustamente nel perseguimento dei propri fini; e, proprio per questo, il problema dell'interpretazione si salda con quello, più radicale, della giustificazione dell'ordine costituzionale, poiché il giudice chiamato a pronunciarsi su questioni costituzionali non si limita a scegliere tra metodi concorrenti, ma è inevitabilmente condotto a confrontarsi con una qualche idea di giustizia, con una certa concezione dei limiti del potere e con il ruolo che l'ordinamento attribuisce alla decisione giudiziale nella tutela dei suoi valori fondamentali. In questa prospettiva, per alcuni il compito del giudice consiste nel lasciarsi orientare dai principi di giustizia che sorreggono l'ordinamento costituzionale, mentre per altri esso implica, più apertamente, la possibilità di prendere in considerazione le politiche pubbliche e le conseguenze delle decisioni, ammettendo che i casi costituzionali possano essere decisi anche alla luce di una valutazione esplicita degli effetti pratici delle diverse soluzioni e dei valori che esse realizzano o sacrificano.

---

[6] Tra i primi contributi di commento e discussione critica della teoria costituzionale di Bobbitt, v. D. HAMILTON, *Rev. of Constitutional Fate: Theory of Constitution*, in *Journal of Policy Analysis and Management*, 2, 4, 1983, 654; D.A. FARBER, *Rev. of Philip Bobbitt: Constitutional Fate*, in *Minnesota Law Review*, 67, 6, 1983, 1329; P.O. GUDRIDGE, Rev. of *False Peace and Constitutional Tradition. Philip Bobbitt: Constitutional Fate*, in *Harvard Law Review*, 96, 8, 1983, 1969; G.R. NICHOL, *Rev. of Giving Substance Its Due, in* P. BOBBITT (a cura di)*, Constitutional Fate;* M. PERRY*: The Constitution, the Courts, and Human Rights*, in *The Yale Law Journal*, 93, 1, 1983, 171; F.P. LEWIS, *Constitutional Fate: Theory of the Constitution, by Philip Bobbitt*, in *The American Political Science Review*, 77, 3, 1983, 751; J. R. SILKENAT, *Rev. of Bobbitt, Philip, Constitutional Fate*, in *The Annals of the American Academy of Political and Social Science*, 474, 1984, 203.

[7] In questo senso, v. M. GOLD, *Rev. of The Rhetoric of Constitutional Argumentation*, in *The University of Toronto Law Journal*, 35, 2, 1985, 154-182.

È precisamente in rapporto a questo duplice scenario che la proposta di Bobbitt acquista il suo rilievo, poiché la sua originalità consiste nel mostrare come molte delle alternative sulle quali si è irrigidito il dibattito (testo o intenzione, metodo o giustizia, interpretazione o policy) siano in realtà false alternative, incapaci di restituire adeguatamente il modo in cui il diritto costituzionale opera nella pratica; sicché il suo contributo non consiste nell'individuazione di un criterio unico del significato costituzionale, ma nel riconoscimento del fatto che il discorso costituzionale si articola attraverso una pluralità di forme argomentative convenzionalmente riconosciute, e che la legittimità stessa del *judicial review* si radica non nell'adesione a un solo metodo, ma nella pratica argomentativa entro cui i giudici effettivamente giustificano le loro decisioni. La sua innovazione non consiste soltanto nell'aver proposto una tipologia delle forme dell'argomentazione costituzionale ma, prima ancora, nell'aver colto la natura retorica del diritto. Come osserva Marc Gold, la sua intuizione più importante, ai fini del dibattito sulla legittimità giudiziaria, consiste proprio nel fatto che «Bobbitt adotta implicitamente una concezione del diritto come forma di retorica (a view of law as a form of rhetoric)»[8].

Secondo Bobbitt, i giudici argomentano ricorrendo a un insieme discreto di forme archetipiche di argomentazione; ciascun tipo di argomento costituisce una convenzione nel senso che rinvia a una plausibile caratterizzazione della Costituzione che trova un radicamento nella pratica interpretativa e che gode di una certa legittimazione nella giurisprudenza. In questa prospettiva, il giudizio sulla correttezza della decisione si colloca sul piano proprio della pratica giuridica e richiede di accertare se, nel caso concreto, la forma di giustificazione adottata sia appropriata, vale a dire se la convenzione argomentativa utilizzata dal giudicante sia pertinente tanto alla funzione esercitata dalla corte quanto ai tratti peculiari della vicenda. Infatti, se nessuna delle modalità argomentative individuate da Bobbitt può essere considerata, in quanto tale «necessariamente illegittima»[9], allora la correttezza della decisione non si misura sulla base di una preferenza teorica astratta per una forma di argomentazione, ma sulla sua congruenza con la pratica[10] in cui essa viene impiegata.

Bobbitt sostiene che la grammatica argomentativa del diritto costituzionale si componga di sei modalità fondamentali: l'argomento storico, testuale, dottrinale, prudenziale, strutturale ed etico. L'argomento storico affronta l'interpretazione costituzionale attraverso la ricerca del significato originario della disposizione da interpretare. L'argomento testuale, invece, non si concentra sulla comprensione originaria, ma sul significato letterale della disposizione in questione. L'argomento dottrinale si incentra sui precedenti, vale a dire sugli standard «che si sono progressivamente accumulati attorno a varie disposizioni costituzionali»[11]. L'argomento prudenziale è guidato dalla preoccupazione per le «circostanze politiche ed

[8] M. GOLD, *op. cit*., 172, trad. nostra. Per Gold, il termine "retorica" si iscrive nella tradizione perelmaniana, richiamandosi espressamente alla definizione di Berman nell'introduzione a *Justice, Law, and Argument*, per cui retorica è «the logic of reasoned discourse, of argumentation, of justification of choices», mettendo così in luce il nesso tra retorica, argomentazione e giustificazione delle decisioni. H. J. BERMAN*, Introduction*, in CH. PERELMAN (a cura di), *Justice, Law, and Argument*. *Essays on Moral and Legal Reasoning*, Dordrecht, 1980, X.

[9] P. BOBBITT, *op. cit*., 139.

[10] Per pratica giuridica si intende, qui, un'attività sociale istituzionalizzata, governata da finalità interne non disponibili alla libera scelta dei partecipanti, che vincolano il senso stesso del giudicare e delle forme della sua giustificazione. In questo senso, F. VIOLA, *Il diritto come pratica sociale*, Milano, 1990, 170: «[u]na pratica esiste se è praticata ed ha una determinata identità se in linea generale è governata da determinate finalità che non dipendono dalla libera scelta dei partecipanti».

[11] P. BOBBITT, *op. cit*., 41 (trad. ns.).

economiche che circondano la decisione»[12], tanto con riguardo alle conseguenze sociali della pronuncia, quanto con riguardo alle ripercussioni della decisione sul ruolo stesso della Corte. L'argomento strutturale procede traendo inferenze «dall'esistenza di strutture costituzionali e dai rapporti che la Costituzione dispone tra tali strutture»[13]. La sesta convenzione è l'argomento etico, che «si fonda su una caratterizzazione (*characterization*) delle istituzioni americane e del ruolo che, al loro interno, spetta al popolo americano. È il carattere (*character*), o *ethos* (*ethos*), della comunità politica americana (*American polity*) a essere richiamato, nell'argomentazione etica, come fonte da cui derivano determinate decisioni»[14]. In questa definizione l'argomento etico viene ricondotto al carattere, o *ethos*, della comunità politica americana. La stessa terminologia, come Bobbitt ammette, non è priva di difficoltà ove precisa che l'argomento etico non va confuso con l'argomento morale in senso proprio, cioè non autorizza il giudice a richiamare una teoria trascendente del bene o del giusto, né gli consentirebbe di sovrapporre immediatamente le proprie preferenze morali all'interpretazione costituzionale; piuttosto rinvia a una concezione pubblica e storicamente determinata del modo in cui una comunità politica comprende se stessa, vale a dire a una certa idea «del tipo di persone che siamo (*the sort of people we are*) e dei mezzi che abbiamo scelto per risolvere i consueti problemi politici e costituzionali (*the means we have chosen to solve political and customary constitutional problems*)»[15]. È rilevante notare che l'argomentazione etica, in questa definizione, è quella cioè che attinge all'immagine che una comunità politica istituzionalmente organizzata ha di sé e del proprio modo di affrontare le questioni fondamentali della convivenza: il richiamo all'etica rinvia anzitutto al *character* di una comunità politica, cioè all'immagine che essa istituzionalmente costruisce di sé, del proprio *ethos* e delle forme ritenute appropriate per affrontare le questioni fondamentali della convivenza[16].

Non sorprende che Bobbitt riconosca il carattere problematico e quasi perturbante di questa modalità argomentativa, ove appunta che: «l'argomento etico è stato trascurato perché è temuto»[17]. E aggiunge che un riconoscimento franco dell'argomento etico è necessario, affinché «altre forme di argomentazione costituzionale (…) non vengano deformate se corrotte per svolgere compiti per i quali sono mal adatte»[18].

---

[12] P. BOBBITT, *op. cit*., 61 (trad. ns.).
[13] P. BOBBITT, *op. cit*., 74 (trad. ns.).
[14] P. BOBBITT, *op. cit*., 94 (trad. ns.).
[15] P. BOBBITT, *op. cit*., 95 (trad. ns.).
[16] In tal senso, riteniamo che l'argomento etico possa essere letto anche alla luce di un lessico prossimo all'etica delle virtù, poiché ciò che viene in rilievo non è soltanto ciò che è consentito fare, ma anche il tipo di soggetto politico che una comunità intende essere e le qualità pratiche che essa riconosce come costitutive della propria autocomprensione. Per una panoramica, sia pure esemplificativa, delle diverse posizioni che hanno segnato la rinnovata attenzione, nel dibattito contemporaneo, per l'etica delle virtù, segnaliamo in particolare G. SAMEK LODOVICI, *Il ritorno delle virtù. Temi salienti delle virtue ethics*, Bologna, 2009; S. VAN HOOFT (a cura di), *The Handbook of Virtue Ethics*, Durham, 2014; *Virtue Ethics: An Overview*, in *Teoria. Rivista di Filosofia*, 38 (2), 2018, 21-31; A. MACINTYRE, *Dopo la virtù. Saggio di teoria morale*, Milano, 1988.
[17] P. BOBBITT, *op. cit*. 137 (trad. ns.). Sullo sfondo emerge il dibattito statunitense sulla *countermajoritarian difficulty*, cioè sulla difficoltà di giustificare, in termini democratici, il potere di giudici non eletti di sindacare decisioni adottate dagli organi rappresentativi. In tale contesto, l'argomento etico è "temuto" perché rende particolarmente visibile il momento valutativo del giudizio costituzionale. Proprio per questo, tuttavia, Bobbitt ritiene che esso non debba essere rimosso: la sua esclusione finirebbe infatti per trasferire impropriamente su altri tipi di argomento compiti giustificativi che essi non sono in grado di svolgere in modo trasparente.
[18] P. BOBBITT, *op. cit*., 167 (trad. ns.).

Il punto è di notevole rilievo: l'argomento etico non può essere espunto senza alterare l'equilibrio complessivo delle forme argomentative, poiché la sua esclusione finirebbe per gravare altre modalità, come quella testuale o giurisprudenziale, di funzioni che esse non sono, da sole, adeguatamente in grado di assolvere.

Secondo Bobbitt, nessuna di queste convenzioni è arbitraria, perché la loro giustificazione non dipende da un criterio puramente soggettivo, ma dal fatto che esse si radicano in premesse endossastiche[19], reputate autorevoli, condivisibili o storicamente accreditate nella cultura costituzionale. La loro natura è, dunque, propriamente retorica: ciascuno di questi argomenti persuade perché richiama elementi che, nel discorso costituzionale, appaiono ragionevoli e riconoscibili[20].

Tali convenzioni costituiscono la «grammatica giuridica»[21] (p. 6) che i giudici sviluppano praticando il diritto costituzionale, così come il parlante di una lingua naturale sviluppa il senso della propria grammatica attraverso l'uso della lingua stessa. La nozione di grammatica è qui particolarmente significativa, perché rinvia ad un ordine pratico e immanente, che si sedimenta nell'esercizio stesso dell'argomentazione: il giudice competente, come il parlante competente, mostra di padroneggiare le forme argomentative nel modo stesso in cui costruisce il proprio discorso.

Ed è precisamente questo aspetto a rendere la nozione di grammatica rilevante per la presente ricerca. Se infatti l'argomentazione giudiziale presenta regolarità sufficientemente stabili da poter essere descritta come una grammatica, allora non è implausibile supporre che un *Large Language Model* possa apprenderne almeno in parte le forme, così come apprende le regolarità di una lingua naturale. Resta però aperta la questione decisiva, che è poi quella al centro di questo lavoro: se tale apprendimento riguardi soltanto la configurazione esterna delle sequenze argomentative, oppure consenta davvero di coglierne il significato giuridico e la funzione giustificativa entro la pratica istituzionale del giudizio.

Come propone Dennis Patterson[22], questo schema, pur essendo nato nel contesto della teoria del *judicial revie*w e del costituzionalismo americano, possiede un valore teorico più ampio e può essere esteso a

[19] E. BERTI, *Gli endoxa in Aristotele e oggi*, in *Endoxa – Prospettive sul presente*, 2(3), 2017, 15-21. «gli *endoxa*, per Aristotele, sono, sì, opinioni, ma non sono opinioni qualsiasi, bensì sono opinioni dotate di un particolare valore, per il fatto di essere condivise, appunto, da tutti, o dalla maggioranza, o dagli esperti, ecc. Aristotele, come è noto, era ottimista dal punto di vista epistemologico, cioè riteneva che, quando tutti la pensano in un certo modo, è molto probabile che essi siano nel vero. C'è infatti un passo dell'Etica Nicomachea in cui egli afferma: «le cose che sembrano a tutti, queste diciamo che sono, mentre chi distrugge questa fiducia, non dirà affatto cose più degne di fede» (X 2, 1173 a 1-2).».

[20] Questa impostazione consente di comprendere anche un altro aspetto decisivo della teoria di Bobbitt: si tratta di un sistema dinamico di reciproche integrazioni e correzioni per cui ogni convenzione non soltanto individua una specifica modalità di giustificazione, ma può essere letta come una risposta alle difficoltà teoriche sollevate dalle altre. Così, ad esempio, l'argomento storico, da un lato, pretende di vincolare il giudice all'intento originario di coloro che hanno creato la Costituzione; tuttavia, esso si espone a limiti evidenti, in particolare all'indeterminatezza e alla manipolabilità delle fonti storiche, oltre che al fatto che, quale che fosse l'intenzione originaria, essa è stata comunque espressa in un testo. Da qui la necessità dell'integrazione con l'argomento testuale. Ma, poiché il significato non risiede intrinsecamente nel testo, non si presenta come oggettivo e non è indipendente da ciò che l'interprete apporta all'atto interpretativo, anche questa convenzione non è autosufficiente. V.P. BOBBITT, *op. cit*. 246.

[21] P. BOBBITT*, op. cit*. 6 (trad. ns.).

[22] D. PATTERSON, *Diritto e verità* (Law and Truth), trad. it. di M. Manzin, Milano, 2010, 207-220.

qualunque decisione giuridica che possa essere giudicata giusta o sbagliata con riferimento al *law sense*[23], cioè al senso della "grammatica giuridica". È appunto questo senso della grammatica giuridica che si forma attraverso l'esercizio delle diverse convenzioni argomentative: la griglia argomentativa consente a ciascun interprete di valutare la correttezza della decisione. Da qui discende anche il rifiuto, da parte di Bobbitt, dell'idea che debba esistere uno e un solo approccio legittimo all'aggiudicazione costituzionale: se l'interpretazione procede mediante argomentazione, e se l'argomentazione è inseparabile dalla persuasione, non è possibile, infatti, stabilire a priori quale tecnica argomentativa debba sempre prevalere. La scelta dell'argomento appropriato dipende dal contesto, dalla valutazione della situazione, dalla individuazione del fine perseguito dalle decisioni giuridiche o dalla giurisprudenza. Per questa ragione, Bobbitt può apparire, secondo la formula di Gold, come «the ultimate Realist»[24] nella misura in cui la legittimità della *judicial review* riposa sul fatto che essa coincide con il modo in cui, nella pratica, i giudici svolgono il proprio compito. Ma, nei suoi stessi termini, si può ritenere che questa teoria riesca effettivamente a legittimare la *judicial review* perché ne mostra il radicamento nella pratica argomentativa che struttura il diritto costituzionale.

L'estensione proposta da Patterson è decisiva per la presente ricerca, perché sposta il fuoco dalla specificità del diritto costituzionale americano alla struttura pratica del ragionare giuridico. Se le modalità bobbittiane non vengono intese come un repertorio legato esclusivamente a uno specifico ordinamento, ma come forme convenzionali di giustificazione attraverso cui una decisione può essere discussa come corretta o scorretta, allora il loro impiego può oltrepassare il contesto originario del *judicial review*[25].

La teoria di Bobbitt apre, per questa nostra ricerca, una pista particolarmente feconda. Se il ragionamento giudiziale si struttura secondo convenzioni argomentative riconoscibili, allora la loro mappatura nelle sentenze può costituire non soltanto un utile strumento di analisi del discorso giudiziario, ma anche il banco di prova per verificare se un *Large Language Model* sia capace di apprenderne la grammatica. In questa prospettiva, la tipologia bobbittiana fornisce una griglia teorica che consente di sottoporre a verifica la capacità di GPT-4o di identificare, raggruppare e classificare le diverse forme della giustificazione giuridica. La questione diventa allora se l'AI riesca a cogliere non soltanto la dimensione testuale della decisione, ma anche la sua forma retorica, ossia la struttura argomentativa attraverso cui il diritto organizza e legittima il discorso del giudice.

3. Il progetto di ricerca: corpus, annotazione e validazione

Su questo sfondo teorico si colloca lo studio condotto nell'ambito di un progetto di ricerca sviluppato nel 2025 da un gruppo interdisciplinare composto da studiosi dell'Università di Trento, della Libera Università di Bolzano e dell'Università di Brescia[26], con l'obiettivo di costruire uno strumento digitale, fondato

[23] La formula *law sense* viene utilizzata in questo contributo per designare non soltanto, come meglio emergerà nelle sezioni conclusive, una sensibilità giuridica (*legal sense*), ma un più radicale *sense of law*, nella misura in cui l'inclusione del sentire nell'esperienza giuridica concorre a ridefinire il concetto stesso di diritto.

[24] M. GOLD, *op. cit*. 175.

[25] Sull'evoluzione dello schema di Bobbit, v. D. PATTERSON, *Diritto e verità*, Torino, 2010, 8.

[26] Più precisamente, il progetto è stato sviluppato da Serena Tomasi, Jacopo Staiano e Carlotta Giacchetta, afferenti all'Università di Trento, da Raffaella Bernardi, afferente alla Libera Università di Bolzano, e da Barbara Montini, afferente all'Università di Brescia. La sua natura interdisciplinare va intesa non solo in senso istituzionale, quale

sull'*argument mining* e sull'impiego di *Large Language Models*, capace di assistere i professionisti del diritto (giudici, avvocati, pubblici ministeri, notai, praticanti) nell'analisi critica delle decisioni giudiziarie considerate nella loro piena complessità argomentativa.

Il punto di partenza del progetto è che la decisione giudiziale non possa essere adeguatamente compresa riducendola né al solo dispositivo, né alla sola massima, né a una semplificazione sillogistica del ragionamento; essa deve invece essere trattata come un prodotto argomentativo stratificato, nel quale la giustificazione della decisione si distribuisce attraverso passaggi molteplici, talora espliciti, talora impliciti, e richiede pertanto strumenti capaci di portarne alla luce la logica interna. In questa prospettiva, il sistema di ausilio digitale non è concepito per sostituire il giurista, ma per assisterlo in un compito cognitivamente oneroso: segmentare il testo, individuare le unità dotate di valore argomentativo e attribuire loro una qualificazione semantica secondo una tipologia teoricamente controllata.

La scelta di assumere come quadro teorico la tipologia di Philip Bobbitt risponde esattamente a questa esigenza. Il progetto non si limita infatti a distinguere tra premesse e conclusioni, né a classificare genericamente i ruoli discorsivi dei segmenti testuali, ma adotta le sei modalità bobbittiane (*historical, textual, structural, prudential, doctrinal, ethical*) come griglia per una classificazione più fine del ragionamento giuridico. A tali categorie è stata aggiunta una categoria residuale, *None*, destinata ai casi in cui non fosse possibile individuare una funzione argomentativa chiara, oppure in cui il segmento risultasse meramente descrittivo o procedurale.

Proprio in questo consiste uno degli elementi più originali del progetto: l'uso di Bobbitt non come semplice schema teorico evocato sullo sfondo, ma come dispositivo operativo per trasformare le decisioni in rappresentazioni argomentative strutturate e, dunque, per verificare se un modello come GPT-4o sia effettivamente in grado di riconoscere la grammatica della giustificazione giudiziale.

Il corpus utilizzato nello studio sperimentale si compone di venti sentenze della Corte di Cassazione, selezionate tra quelle indicate come più rilevanti sul sito ufficiale della Corte, nella sezione "Novità". Il corpus comprende dieci decisioni civili e dieci decisioni penali; tutte le sentenze sono in lingua italiana e sono state estratte dalla banca dati *De Jure*[27]. Le decisioni civili coprono l'arco temporale 2018-2025, mentre quelle penali si concentrano sugli anni 2023-2024. Anche sotto il profilo materiale il corpus è volutamente eterogeneo: la lunghezza dei provvedimenti varia da decisioni relativamente concise, di circa quattro pagine, a pronunce assai più estese, fino a ventisei pagine. Tale varietà è metodologicamente rilevante, perché consente di mettere alla prova sia l'annotazione umana sia quella automatica su testi non uniformi per densità argomentativa, struttura e stile redazionale.

Prima dell'annotazione, tutti i testi sono stati sottoposti a una fase di pre-processing, volta a estrarre il testo integrale ed eliminare le parti non argomentative, come intestazioni, metadati o sommari procedurali.

---

collaborazione tra studiosi appartenenti a più atenei, ma anche in senso epistemico e metodologico, poiché la ricerca si colloca all'intersezione fra teoria dell'argomentazione, ermeneutica giuridica, *digital humanities*, intelligenza artificiale e linguistica computazionale. Il progetto ha costituito altresì oggetto dell'elaborato finale di Barbara Montini, intitolato *AI e Dominio Giuridico: Strumenti Computazionali per la Giustizia e la Pubblica Amministrazione*, redatto nell'ambito del Master interuniversitario di II livello in *Intelligenza Artificiale, Mente, Impresa*, con relatrice Raffaella Bernardi e correlatrice Serena Tomasi, nell'anno accademico 2023/2024; in tale lavoro l'autrice dà conto della propria partecipazione al progetto nell'ambito dello stage del Master e ne ricostruisce presupposti, metodo e finalità.

[27] Banca dati De Jure, https://dejure.lefebvregiuffre.it/ (ultima consultazione 01/03/2026).

La pipeline di annotazione si articola in una sequenza di passaggi chiaramente distinti. Essa non va intesa come una generica interazione con il modello, ma come una procedura definita, composta da conversione dei file PDF in una struttura XML, classificazione dei paragrafi secondo la loro funzione argomentativa, raggruppamento semantico, categorizzazione bobbittiana, reintegrazione delle annotazioni nel file XML ed esportazione finale in Excel. In termini operativi, i file PDF delle sentenze venivano pre-processati per estrarre il testo integrale e rimuovere le sezioni non argomentative; il testo così ottenuto veniva quindi segmentato in paragrafi, ciascuno racchiuso in un tag dedicato e dotato di un identificatore univoco. Successivamente, ogni paragrafo veniva etichettato dal modello come *premise*, *conclusion* oppure *null*; in una fase distinta, i paragrafi venivano raggruppati semanticamente e i gruppi così ottenuti venivano classificati secondo le categorie bobbittiane. Le annotazioni, inizialmente prodotte in formato JSON, venivano poi reintegrate nell'XML e infine esportate in Excel, così da renderle più facilmente esaminabili dagli annotatori umani.

La prima fase sostanziale della pipeline è la segmentazione testuale. In questa fase, GPT-4o veniva interrogato mediante un prompt strutturato che gli richiedeva di dividere il testo in paragrafi coerenti e di etichettare ciascun paragrafo come *premise*, *conclusion* oppure *null*. A ogni paragrafo veniva inoltre attribuito un identificatore costruito secondo la logica delle catene argomentative: una lettera indicava la catena di riferimento, mentre un numero progressivo segnalava l'ordine del segmento all'interno di quella catena. Nei prompt, la catena argomentativa era definita come un argomento a sostegno della conclusione finale relativa a uno specifico motivo di impugnazione, insieme alle controargomentazioni considerate dalla Corte. Poiché il testo veniva processato in più *chunk* a causa dei limiti di lunghezza del modello, la pipeline prevedeva un messaggio dinamico volto a garantire la continuità della numerazione delle catene tra un blocco e l'altro. Questo aspetto è metodologicamente importante, perché mostra che la segmentazione non è trattata come mera suddivisione grafica, ma come un primo tentativo di rendere leggibile il flusso del ragionamento giudiziale.

La seconda fase è quella del raggruppamento semantico. Una volta ottenuti i paragrafi e i loro identificatori, GPT-4o veniva interrogato mediante un prompt distinto, volto a raggruppare le unità che condividevano una medesima logica semantica o affrontavano lo stesso tema. In questa fase, l'ordine dei paragrafi e i loro identificatori non dovevano guidare la decisione: il raggruppamento doveva fondarsi esclusivamente sul significato. Ogni gruppo poteva comprendere fino a sette/otto paragrafi; le unità prive di sufficiente connessione tematica restavano non raggruppate e ricevevano l'indicazione *group_id: null*. Anche in questa fase, la continuità tra *chunk* successivi veniva mantenuta tramite un messaggio dinamico nei prompt, volto a impedire riassegnazioni arbitrarie dei gruppi già creati.

Solo a questo punto interviene la terza fase, quella della categorizzazione vera e propria. Ciascun gruppo semantico veniva passato a GPT-4o, al quale era richiesto di selezionare la categoria bobbittiana più appropriata tra quelle predefinite, oppure di assegnare l'etichetta *None* nei casi in cui nessuna delle sei modalità risultasse adeguata, accompagnando tale scelta con una breve spiegazione. Sia il raggruppamento sia la categorizzazione sono stati condotti mediante prompt zero-shot con istruzioni strutturate e temperatura pari a 0.2. Al termine del processo, ciascun paragrafo risultava annotato nell'XML con almeno tre informazioni: l'*id*, il *group_id* e la categoria assegnata.

Nel prompt di classificazione, le categorie vengono descritte in modo sintetico ma operativo: l'argomento storico come ricorso alle intenzioni originarie dei *framers* e dei *ratifiers*; quello testuale come fondato sul

significato letterale delle parole; quello strutturale come analisi del sistema costituzionale nel suo complesso; quello prudenziale come valutazione di pro e contro pratici e delle conseguenze sociali; quello dottrinale/giurisprudenziale come uso dei precedenti; quello etico come appello all'*ethos*.

Una componente essenziale del progetto è costituita dalla validazione umana. Per controllare l'output della pipeline automatica, gli autori hanno raccolto annotazioni manuali da parte di cinque esperti con diverso livello di formazione giuridica: tra i junior figurano uno studente di giurisprudenza, due praticanti forensi e un dottorando; il polo senior è rappresentato da un professore universitario del settore giuridico. A ciascun annotatore è stato fornito un file Excel strutturato, contenente i paragrafi già raggruppati e dotati di *group_id*; per ogni paragrafo, l'annotatore doveva selezionare tramite menu a tendina la categoria bobbittiana ritenuta più appropriata. Per assicurare la comparabilità con il comportamento del modello, gli annotatori hanno ricevuto gli stessi prompt impiegati nella fase automatica. Inoltre, qualora un gruppo semantico apparisse incoerente o internamente disomogeneo, l'annotatore era comunque tenuto ad assegnare la categoria ritenuta prevalente, ma poteva contrassegnare quel gruppo come *incorrect* e proporre un raggruppamento alternativo. Questo protocollo mostra che l'annotazione non costituisce un compito meccanico, ma è un'attività interpretativa riflessiva, capace di produrre anche un riscontro qualitativo sulla bontà del raggruppamento generato dal modello. Gli annotatori junior hanno esaminato solo una parte del corpus, mentre l'annotatore senior ha annotato l'intero insieme delle sentenze; il tempo richiesto per ciascuna decisione è variato da 30 a 120 minuti, segno della difficoltà intrinseca del compito. Il protocollo adottato non costruisce dunque un *gold standard* forte in senso stretto, ma una base di confronto utile a fini esplorativi, coerente con la finalità del lavoro ma non ancora sufficiente, da sola, a fungere da riferimento umano pienamente consolidato.

La fase di valutazione è stata impostata in modo da tener conto della pluralità interpretativa tipica del ragionamento giuridico. Per misurare la coerenza delle annotazioni del modello rispetto a quelle umane, i ricercatori hanno adottato due strategie complementari, entrambe fondate su Cohen's kappa. Nella valutazione *intersection-based*, si considera soltanto il sottoinsieme dei casi in cui due annotatori umani abbiano attribuito indipendentemente la stessa etichetta al medesimo paragrafo: questo sottoinsieme viene trattato come *gold standard* ad alta affidabilità, e l'output di GPT viene confrontato con esso. Nella valutazione *union-based*, invece, il modello è considerato corretto se la sua classificazione coincide con quella di almeno uno dei due annotatori. Questa seconda modalità è più permissiva, ma anche più vicina alla realtà dell'interpretazione giuridica, perché riconosce che in testi complessi e ambigui possono esistere più attribuzioni plausibili. La combinazione delle due misure consente di osservare il comportamento del modello sia in condizioni di forte convergenza umana, sia in un quadro più realistico di disaccordo interpretativo.

I risultati quantitativi confermano la difficoltà intrinseca del compito. L'accordo tra annotatore senior e junior è modesto: 0.17 nel settore penale e 0.27 in quello civile. L'accordo tra senior e GPT è pari a 0.15 nel penale, ma diventa negativo (-0.03) nel civile; quello tra junior e GPT è 0.07 nel penale e -0.09 nel civile. Se si guarda alla misura *intersection-based* senior∩junior vs GPT, il dato resta 0.15 nel penale e scende a -0.0936 nel civile; la misura *union-based* migliora, arrivando a 0.46 nel penale e 0.1874 nel civile. A livello di classi, il modello ottiene i risultati migliori sulla categoria *None* e, in misura minore, sulla categoria *Doctrinal*: nel penale, ad esempio, la precisione/recall/F1 per *Doctrinal* è 0.41 / 0.80 / 0.54, mentre

per *None* è 0.95 / 0.87 / 0.91; nel civile, *Doctrinal* registra 0.39 / 0.76 / 0.52, *None* 0.77 / 0.45 / 0.57, *Textual* 0.50 / 0.18 / 0.27, mentre *Ethical* ha precision, recall e F1 pari a 0.00 / 0.00 / 0.00.

Si evince che GPT-4o tende a catturare solo un sottoinsieme limitato delle convenzioni argomentative (testuale, dottrinale in via prevalente), mancando di cogliere le sfumature semantiche e contestuali che gli esperti legali riescono a individuare grazie alla conoscenza del dominio; il comportamento del modello appare, nel complesso, più vicino a quello di un *junior expert* che a quello dell'annotatore senior.

Il basso accordo inter-annotatore conferma la difficoltà intrinseca del compito, ma suggerisce al tempo stesso cautela nell'interpretazione dei risultati del modello. Se il benchmark umano è solo parzialmente consolidato, anche la valutazione dell'output di GPT-4o non può essere considerata definitiva. I risultati discussi vanno dunque intesi come evidenza esplorativa e teoricamente significativa, ma non ancora come una verifica conclusiva.

Considerato nel suo insieme, l'esperimento non si lascia ridurre a un semplice test di accuratezza classificatoria di un LLM, ma costituisce una verifica empirica di una questione teorica più ampia, cioè se un *general-purpose LLM*, quale GPT-4o, possa essere impiegato come strumento di pre-interpretazione strutturata del testo giudiziario.

L'evidenza raccolta appare coerente con l'ipotesi che il modello possa offrire un aiuto reale nel pre-processing, nella segmentazione e in una prima organizzazione del materiale argomentativo, ma mostra anche che il suo rendimento resta tuttavia limitato proprio là dove la ricostruzione del ragionamento esige sensibilità al contesto giuridico e alla differenza tra forme concorrenti di giustificazione.

4. L'argomentazione etica come punto di rottura

Il dato più significativo dello studio sperimentale non è semplicemente che GPT-4o ottenga risultati diseguali, ma che il suo rendimento peggiori proprio quando il compito richiede di riconoscere forme di giustificazione meno frequenti, meno formulari e più cariche di implicazioni normative. In questo senso, la categoria dell'argomentazione etica costituisce il banco di prova più istruttivo.

Nel sottoinsieme civilistico, essa compare una sola volta nel test set umano, ma in quel caso il modello registra precisione, recall e F1 pari a zero: non si tratta dunque di una semplice performance bassa, bensì dell'incapacità di intercettare quella modalità argomentativa nel momento stesso in cui compare. Il dato non può essere liquidato come irrilevante solo perché la classe è rara; al contrario, proprio la rarità della categoria consente di vedere con particolare nettezza il limite strutturale del modello, che tende a riconoscere soprattutto ciò che è statisticamente saliente e linguisticamente più marcato, mentre fallisce quando la giustificazione si affida a passaggi più sottili, meno stereotipati e più dipendenti dal contesto istituzionale.

Questo fallimento è tanto più importante se si considera che, nel framework adottato, l'argomento etico non coincide con un generico richiamo a valori, ma designa una modalità specifica di giustificazione, fondata su principi morali o ideali condivisi, e perciò richiede di distinguere quando un passaggio della decisione non si limita a descrivere un bene giuridico o a evocare una finalità normativa, ma mobilita quel bene o quella finalità come ragione decisiva della conclusione. Il problema, allora, non è soltanto lessicale. Un modello che lavora per regolarità linguistiche può riconoscere facilmente citazioni, formule tipiche, richiami giurisprudenziali consolidati; incontra, invece, difficoltà laddove deve cogliere il momento in cui

il testo passa dal piano della descrizione o della sistemazione dottrinale o giurisprudenziale a quello della giustificazione assiologica.
Il risultato zero sulla categoria etica segnala precisamente questa soglia: il modello può riprodurre il linguaggio dei valori, ma non per questo riconosce quando esso opera, nel ragionamento giudiziale, come forma autonoma di giustificazione.
Non è casuale, del resto, che GPT-4o si comporti, nel complesso, in modo più simile a un *junior expert* che a un annotatore senior, mostrando una familiarità di base con alcune distinzioni argomentative, ma non quella profondità necessaria per intercettare categorie meno frequenti o più concettualmente esigenti, tra le quali quelle *Ethical* e *Prudential*. Il punto, dunque, non è che il modello sbagli occasionalmente un'etichetta, ma che esso tenda a perdere di vista proprio le modalità del ragionamento giuridico che più si discostano dalla dimensione "formularia" della decisione e che esigono una comprensione più fine della sua funzione giustificativa.

5. I limiti del riconoscimento della grammatica argomentativa

Un altro risultato da mettere a fuoco riguarda la pretesa, almeno implicita, che il modello sia capace di apprendere la grammatica argomentativa del diritto.
Se si guarda ai dati con attenzione[28], ciò che emerge non è il riconoscimento affidabile di una grammatica, ma piuttosto la capacità di cogliere alcune regolarità locali del discorso giudiziario. La stessa valutazione d'insieme mostra un quadro fragile: l'accordo tra annotatore senior e junior è già modesto (0.17 nel penale e 0.27 nel civile) e questo conferma la complessità intrinseca del compito; ma l'accordo tra senior e GPT-4o si ferma a 0.15 nel penale e scende addirittura a -0.03 nel civile, mentre quello tra junior e GPT-4o è 0.07 nel penale e -0.09 nel civile. Anche quando si adotta la misura più permissiva, fondata sull'unione delle annotazioni umane, il valore resta molto diverso nei due settori: 0.46 nel penale e 0.1874 nel civile. Nel complesso, questi numeri non descrivono la padronanza di una grammatica, ma una corrispondenza intermittente, che si mantiene solo in parte e solo in alcune condizioni.
La distribuzione per categorie conferma questa impressione. Il modello ottiene infatti i risultati migliori su *None*, cioè sulla classe residuale dei passaggi descrittivi o procedurali, e risultati discreti su *Doctrinal*, la categoria che più facilmente si lascia riconoscere attraverso indizi esteriori come i richiami ai precedenti o a linee giurisprudenziali consolidate. Nel penale, *Doctrinal* registra precisione 0.41, recall 0.80 e F1 0.54, mentre *None* arriva a 0.95 / 0.87 / 0.91; nel civile, *Doctrinal* si colloca a 0.39 / 0.76 / 0.52, *None* a 0.77 / 0.45 / 0.57, e *Textual* a 0.50 / 0.18 / 0.27. Per contro, nel penale *Prudential, Structural e Textual* registrano tutte valori nulli, e nel civile la categoria *Ethical* è interamente mancata. Il quadro che se ne ricava è chiaro: il modello intercetta ciò che è più visibile, frequente e linguisticamente stereotipato, ma non riesce a distribuire correttamente l'intero spettro delle modalità argomentative.
Sotto questo profilo, si può dire che GPT-4o non apprende davvero la grammatica degli argomenti, bensì alcune delle sue tracce superficiali. Una grammatica, in senso forte, non è la mera reiterazione di pattern

[28] Il report dei risultati è disponibile in C. GIACCHETTA, R. BERNARDI, B. MONTINI, J. STAIANO, S. TOMASI, *Argumentative Analysis of Legal Rulings: A Structured Framework Using Bobbitt's Typology*, in E. CHISTOVA, P. CIMIANO, S. HADDADAN, G. LAPESA, R. RUIZ-DOLZ (a cura di), *Proceedings of the 12th Argument Mining Workshop*, Vienna, 2025, 107-115, consultabile all'indirizzo https://aclanthology.org/2025.argmining-1.10/ (ultima consultazione 01/03/2026).

ricorrenti, ma l'insieme delle regole di pertinenza che consente di distinguere che cosa, in un determinato contesto, conta come argomento testuale, che cosa conta come argomento strutturale, che cosa conta come argomento etico.

Ora, il protocollo sperimentale chiedeva precisamente questo: segmentare il testo in unità argomentative, raggrupparle secondo coerenza semantica e attribuire ai gruppi una modalità di giustificazione. Ma i risultati mostrano che il modello non riesce a stabilizzare tale operazione: la sua classificazione non si distribuisce sull'intera grammatica bobbittiana; al contrario, si concentra su poche classi più facili e lascia scoperte proprio quelle che richiedono un passaggio dall'analisi semantica al riconoscimento della funzione giustificativa.

Il problema, invero, non dipende soltanto dal modello, ma anche dalla difficoltà del compito. Gli stessi autori sottolineano che la classificazione dei testi giudiziari è intrinsecamente complessa e soggettiva, e che perfino gli annotatori junior faticano a raggiungere pieno accordo. Proprio per questo, la comprensione della sentenza non può essere ridotta né a un'attività di sintesi né a un mero rilevamento di ricorrenze linguistiche: essa implica, piuttosto, un'operazione interpretativa più esigente, che richiede quel *law sense*[29] necessario a riconoscere la struttura argomentativa e la funzione giustificativa della decisione.

6. Dal *legal sense* al *law sense*: i limiti degli LLM nel giudizio

Le conclusioni dell'analisi richiedono di esplicitare una distinzione che ne sorregge l'intero impianto, quella tra *legal sense* e *law sense*. Con la prima espressione si può designare la capacità di riconoscere, organizzare e trattare materiali giuridici secondo regolarità linguistiche, testuali e sistematiche: segmentare una decisione, individuare ricorrenze, raggruppare passaggi semanticamente affini, formulare ipotesi classificatorie. Con *law sense*, invece, si intende una capacità ulteriore, quella di "sentire" la pertinenza delle ragioni, il loro peso relativo, la loro adeguatezza rispetto al caso concreto e al contesto istituzionale in cui la decisione si colloca. È precisamente su questo secondo terreno che si manifesta il limite strutturale degli LLM: essi possono mostrare un certo *legal sense* operativo, ma non dispongono ancora di quel *law sense* che consente di cogliere la grammatica viva della giustificazione giuridica. In questa prospettiva, il *law sense* può essere inteso nel solco della riflessione di M. Paola Mittica sul «senso del sentire»[30], là dove la ricerca del senso dell'esperienza giuridica si apre nello spazio dell'alterità che attribuisce al diritto il ruolo di strumento di comprensione e tutela delle relazioni umane, irriducibile per questo ad una tecnica impersonale o ad una procedura burocratica.

[29] Sulla c.d. *sensory turn* e sul rilievo della dimensione sensoriale ai fini di una comprensione integrale dell'esperienza giuridica, v. D. Mandić, C. Nirta, A. Pavoni, A. Philippopoulos-Mihalopoulos (a cura di), *Law and the Senses Series: The Taste Issue*, in *Non Liquet: The Westminster Online Working Papers Series*, 2013; D. Mandić, C. Nirta, A. Pavoni, A. Philippopoulos-Mihalopoulos (a cura di), *Law and the Senses Series: The Smell Issue*, in *Non Liquet: The Westminster Online Working Papers Series,* 2015; D. Mandić, C. Nirta, A. Pavoni, A. Philippopoulos-Mihalopoulos (a cura di), *Law and the Senses Series: The Touch Issue*, in *Non Liquet*, London, 2016; D. Mandić, C. Nirta, A. Pavoni, A. Philippopoulos-Mihalopoulos (a cura di), *Law and the Senses Series: The See Issue*, in *Non Liquet*, London, 2018. La serie mostra, in particolare, come il diritto non si esaurisca nella sola dimensione testuale o concettuale, ma si dia anche come esperienza sensibile, radicata nella vita concreta, onde la sensibilità non può essere espunta dalla ricerca della razionalità giuridica.

[30] M.P. Mittica, *Senso del sentire. Law and Humanities ed Estetica giuridica*, in *Rivista di filosofia del diritto*, 441–456.

La pipeline mostra certamente una utilità concreta e non trascurabile nella segmentazione analitica della sentenza in premesse, conclusioni e plessi argomentativi. Si tratta di un'operazione tutt'altro che elementare: isolare i passaggi che svolgono una funzione giustificativa, distinguerli da quelli meramente descrittivi o ricostruttivi, e riconoscere i nuclei discorsivi entro cui si sviluppa una determinata linea di ragionamento richiede un lavoro interpretativo attento, lento e cognitivamente oneroso. Lo conferma il fatto che gli annotatori umani hanno svolto questa attività impiegando un tempo assai considerevole, variabile, a seconda della complessità della decisione, tra trenta e centoventi minuti per singola sentenza.

Sotto questo profilo, il vantaggio offerto dal sistema è evidente: consente di eseguire in tempi rapidi una prima articolazione strutturata del testo, rendendo immediatamente visibili i segmenti potenzialmente rilevanti, le sequenze di ragionamento e i principali argomenti. Anche quando tale articolazione richieda di essere successivamente verificata, corretta o raffinata dall'interprete umano, costituisce comunque un supporto metodologico prezioso, perché abbrevia in modo significativo la fase preliminare di lettura e di disposizione argomentativa. In altri termini, il sistema non sostituisce il giudizio del giurista, ma può alleggerire il costo cognitivo della ricostruzione argomentativa, soprattutto quando ci si confronta con testi lunghi, densi e internamente stratificati. Ed è precisamente in questa funzione di pre-strutturazione che oggi si colloca uno dei vantaggi più chiari dell'impiego degli LLM nell'analisi delle decisioni giudiziarie.

I dati mostrano però un limite prestazionale di GPT-4o, rilevando che la riconoscibilità linguistica di alcune forme argomentative non coincide con la ricostruzione della grammatica retorica del diritto. Laddove l'argomento è dottrinale/giurisprudenziale o dove il testo è semplicemente non argomentativo, il modello può fornire un aiuto plausibile; ma quando la decisione si organizza attraverso modalità meno frequenti, più contestuali o più assiologicamente dense, la prestazione si incrina. L'argomento etico, in particolare, rappresenta il luogo in cui questa incrinatura diventa visibile in forma quasi paradigmatica, perché lì il diritto non si limita a combinare fonti, precedenti o strutture, ma prende posizione attraverso una giustificazione che implica il rilievo normativo dei valori. In questo passaggio, il modello non riconosce più la grammatica dell'argomentazione.

I risultati dell'esperimento non autorizzano una conclusione negativa: non mostrano che i *Large Language Models* siano inutilizzabili nel lavoro giuridico, ma che la loro utilità deve essere collocata al livello corretto. Allo stato attuale, un sistema come GPT-4o è certamente in grado di svolgere con una qualche efficacia compiti preliminari e ausiliari: può riassumere decisioni anche complesse, individuare ricorrenze lessicali e tematiche, proporre una prima segmentazione del testo, raggruppare passaggi semanticamente vicini e, almeno nei casi più visibili e formulari, suggerire ipotesi di classificazione argomentativa. Il suo contributo, tuttavia, resta confinato a ciò che potremmo chiamare il livello della plausibilità organizzativa del discorso: esso ordina, connette, rende salienti alcune strutture, ma non per questo accede ancora al senso propriamente giuridico della giustificazione.

In ciò emerge quello che riteniamo sia il limite decisivo. Ciò che manca al modello non è soltanto una competenza più raffinata nella classificazione delle categorie rare, né una migliore sensibilità alla variabilità semantica dei testi; ciò che manca è, più radicalmente, il *law sense*[31], vale a dire quel senso del diritto che non coincide con la mera conoscenza di formule, precedenti o configurazioni linguistiche, ma rinvia a una capacità di sentire la pertinenza delle ragioni, il loro peso, la loro appropriatezza rispetto al caso, al

[31] Per una definizione di «senso» e delle implicazioni del sentire, ci riferiamo in particolare a M.P. MITTICA, *Senso del sentire. Law and Humanities ed Estetica giuridica*, in *Rivista di filosofia del diritto***,** VIII, 2, 2019, 441-456.

contesto istituzionale e alla posizione dei soggetti coinvolti. Valorizzare la dimensione sensibile implica riconoscere che il diritto non è soltanto un sistema di testi né una rete di inferenze, ma è anche una pratica di giudizio che richiede un orientamento sensibile verso ciò che, in una determinata situazione, conta come ragione giuridicamente seria, come giustificazione adeguata, come equilibrio accettabile tra istanze concorrenti. Il giurista non si limita mai a registrare occorrenze o a distribuire enunciati entro una griglia formale; esercita una competenza che è insieme concettuale, pratica e percettiva, perché deve cogliere ciò che nel caso rileva, ciò che nel testo pesa, ciò che nella giustificazione persuade non solo perché è ben formato, ma perché è internamente adeguato alla pratica del diritto. Questo "sentire giuridico" non ha nulla di irrazionale o di meramente soggettivo: esso designa piuttosto quella familiarità disciplinata con la grammatica viva del diritto che consente di distinguere tra un argomento soltanto possibile e un argomento realmente pertinente, tra una formulazione astrattamente corretta e una ragione che, nel contesto della decisione, può essere assunta e difesa come propria.

Da questo punto di vista, il limite dell'LLM non consiste soltanto nel fatto che esso non "capisce" come un essere umano, ma nel fatto che non partecipa a quella forma di esperienza istituzionale entro cui le ragioni acquistano il loro senso giuridico. Il modello può riconoscere pattern, simulare la forma esterna della giustificazione, riprodurre il linguaggio dei valori e persino anticipare alcune connessioni tra segmenti del testo; ma non dispone di quel *sentire* che lega il giudizio alla responsabilità, alla misura, alla rilevanza del contesto, alla coscienza dei limiti. In altri termini, gli manca quella specifica sensibilità giuridica (*law sense*) che consente di avvertire quando un argomento è davvero decisivo, quando è accessorio, quando è forzato, quando è elusivo, quando copre con veste tecnica una scelta assiologica, e quando invece esprime una presa di posizione che l'ordinamento può riconoscere come giustificata.

Va messo in evidenza che i risultati dell'esperimento devono essere letti anche alla luce dei limiti del protocollo di validazione adottato: se il basso accordo tra annotatori conferma la difficoltà intrinseca del compito, esso suggerisce al tempo stesso la necessità di un ulteriore rafforzamento metodologico; in questo senso, una prosecuzione della ricerca richiederebbe un numero maggiore di annotatori sulle medesime sentenze o sui medesimi gruppi semantici, un momento di *adjudication* volto alla costruzione di un benchmark condiviso e una verifica preliminare della stabilità delle categorie bobbittiane prima ancora della valutazione del modello. In tal senso, i risultati qui discussi appaiono teoricamente significativi, ma vanno intesi con la cautela imposta da un riferimento umano che, allo stato, non può ancora dirsi pienamente consolidato.

Per questa ragione, il risultato più importante della ricerca non è la semplice constatazione che GPT-4o abbia difficoltà nel riconoscimento dell'argomentazione etica, o che non ricostruisca in modo sufficientemente stabile la grammatica degli argomenti; è piuttosto la messa in evidenza del fatto che tra regolarità linguistica e senso giuridico permane uno scarto che non può essere colmato con il solo aumento della potenza computazionale o con una migliore ingegneria del prompt. Gli LLM possono certamente essere utili, e lo sono già, come strumenti di supporto all'organizzazione del materiale normativo e giurisprudenziale; ma il passaggio dalla ricorrenza al significato, dalla connessione semantica alla giustificazione, resta affidato a quella facoltà di giudizio che appartiene a soggetti inseriti in una pratica istituzionale e dotati di *law sense*, «consapevoli del fatto che il "senso" non si esaurisce nelle forme chiuse dei codici: ché è "senso" anche il "sentire" e il "sentimento"»[32].

[32] M.P. MITTICA, *op. cit.*, 452.

Proprio per questo, crediamo che il loro uso in ambito giudiziario debba essere pensato non in termini sostitutivi, ma in termini rigorosamente ausiliari: come strumenti che possono sostenere il lavoro dell'interprete, ma non assumere il posto di quel sentire giuridico attraverso cui il diritto, propriamente, si comprende e si applica.