A path to legal coherence through AI

Steve Huntsman*, Michael Robinson**, and Ludmilla Huntsman***

* sch213@nyu.edu; ORCID:0000-0002-9168-2216
** michaelr@american.edu; ORCID:0000-0003-0766-3301
*** ludmilla@cogsecalliance.org; ORCID:0009-0002-6599-0941

Coherence is rare in life. Law, administration, and jurisprudence are riddled with inconsistencies.[1] Disinformation, doublethink, hypocrisy, and bullshit are ubiquitous. However, there may be a technological remedy (we do not say cure). In this article, we i) quantitatively illustrate incoherence in immigration courts and frame coherence in the context of law; ii) provide evidence that large language models (LLMs) can accurately compile local information into natural data structures that enable coherence-driven inference; iii) outline how efficient computation of globally coherent substructures enables useful forms of machine cognition; and iv) discuss how the technology we describe can most plausibly be developed and deployed with an eye towards legal and governmental applications.

Asylum decisions are quantitatively incoherent

The asylum system is well suited as a natural experiment for gauging whether or how laws are applied consistently and coherently. First, caseloads are large and *a priori* statistically fairly regular within a given immigration court. Second, outcomes are well approximated by a simple binary characterization: asylum is usually either denied or granted. Third, there are readily accessible data organized by the Transactional Records Access Clearinghouse (TRAC) at Syracuse University, which points out that:

> *Within a single Court when cases are randomly assigned to judges sitting on that Court, each Judge should have roughly a similar composition of cases given a sufficient number of asylum cases.*

In fact, asylum decisions and caseloads often vary widely across judges in most immigration courts, as Figure 1 illustrates for the two largest immigration courts (as measured either by the numbers of judges or of asylum cases) in New York and San Francisco.

[1] We use the word "consistency" to refer to how data fit together "locally" and the word "coherence" to refer to how data fit together "globally." A similar distinction has been drawn by Luc Wintgens in work on "legisprudence."

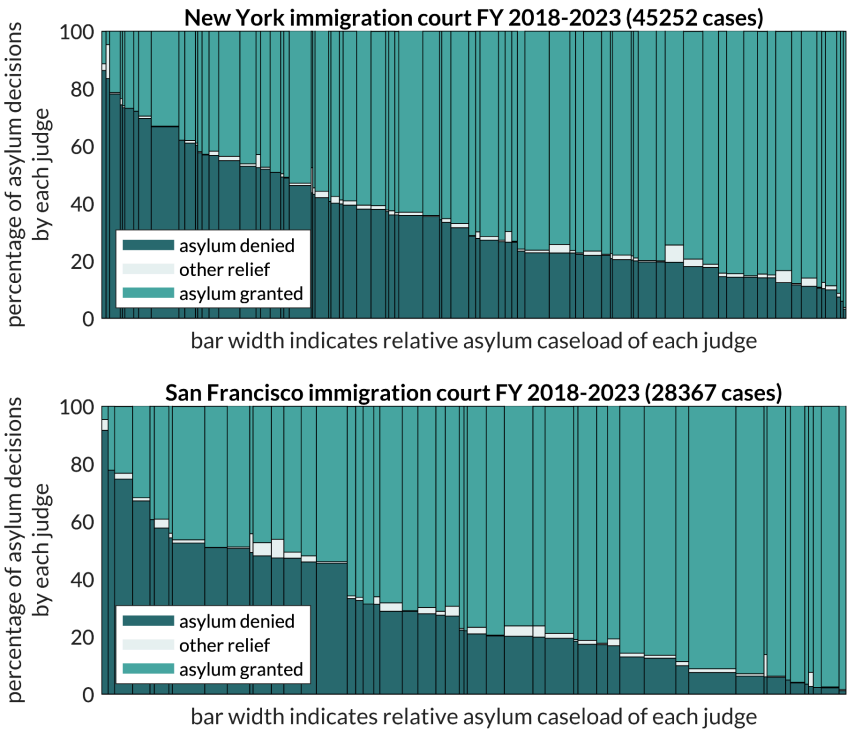


Figure 1. Asylum decisions and caseloads often vary widely across judges within a given court. Some other courts exhibit considerably less slope in plots such as this, reflecting more equitable outcomes and presumably more equal treatment. If this were not the case, a more fine-grained statistical analysis might be warranted to reach any conclusions.

The sloping bar charts in Figure 1 give cause for deep concern regarding a core tenet of justice – and prerequisite for any substantive notion of legal coherence – to "treat like cases alike." But TRAC offers a lifeline that parallels an argument for dispensing with consistent treatment:

> *…variations in asylum decisions among Judges on the same Immigration Court would appear to reflect, at least in part, the judicial philosophy that the Judge brings to the bench. However, if judges within a Court are assigned to specialized dockets or hearing locations, then case compositions are likely to continue to differ and can contribute to differences in asylum denial rates.*

However, a bit of data analysis suggests that immigration judges collectively do not treat like cases alike, and that judicial philosophies and immigration law are applied incoherently even within individual courts.

Comparing the distributions of asylum seeker nationalities (inputs) and asylum decisions (outputs) across judges in the same courts during FY 2018-2023 exposes some startling divergences. For example, this pair of judges in the Arlington immigration court had very similar input distributions and very different output distributions: one denied asylum in over 93% of cases, and the other denied asylum in less than 40 percent of cases. Both judges decided between 270 and 300 cases

during the period. It is not credible to assume that their asylum decisions and underlying judicial philosophies are mutually consistent.

While our analysis has indicated that there are many instances of judges in the same courts apparently deciding cases with a modicum of pairwise consistency, it has also uncovered many other instances where very similar caseloads lead to remarkably different decisions. For example, this pair of judges in Baltimore, or this pair in Boston, or these three judges in Houston, or this pair or this pair or this pair in Los Angeles, or this pair in Miami, or this pair or this pair or this pair in San Francisco, all make very different decisions on similar caseloads.

In short, the evidence is clear that asylum decisions are often incoherent, arbitrary, and capricious at the court level, even if individual judges are usually (self-) consistent.

Law is incoherent

If there is any surprise here, it is only the granular quantification of something that has long been commonly understood. Over 50 years ago, Marvin Frankel's book *Criminal Sentences: Law Without Order* spurred the development of sentencing guidelines by pointing out that:

> *the evidence is conclusive that judges of widely varying attitudes on sentencing, administering statutes that confer huge measures of discretion, mete out widely divergent sentences where the divergences are explainable only by the variations among the judges, not by material differences in the defendants or their crimes.*

Based on the quantitative evidence that asylum cases are often decided incoherently, it is hard to imagine that the judicial branch proper does not fare worse. Indeed, recent evidence indicates that sentencing guidelines are also applied incoherently. Moreover, since the Supreme Court overturned *Chevron* deference in *Loper Bright Enterprises v. Raimondo*, it also seems likely that jurisprudential incoherence in administrative law will metastasize. Though this sea change may reduce *temporal* inconsistency, in our view this is a dubious goal over sufficiently long timescales, since policies and interpretations must evolve over time to be responsive to circumstances. In any event, the goal of reducing temporal inconsistency may be thwarted anyway by tension between the surviving applicable doctrines.

At least one mathematical model shows how incoherence can naturally arise via multi-step reasoning using abstract subjective criteria that depend on objective facts. At the same time, judges seem to behave more incoherently than juries, and there is an obvious explanation: a sample size of one admits more variation than a sample size of twelve.

This state of affairs has motivated inquiries into the nature of coherence in law and also – perhaps out of a sense of defeatism – substantive defenses of incoherence in law. Arguments from pluralism, ambiguity, and discretion have variously been advanced in this direction, with

prominent themes that equality before the law need not amount to equity among outcomes or that laws must be applied consistently over time (see, for instance, here, here, here, and here).

Coherence is a worthy goal

In our view, all of the arguments defending legal incoherence are hollow nods to expediency at the expense of fairness. Arguments against *res judicata* or estoppel seem less likely than those against coherence precisely because the practical grounding of the former doctrines makes them expedient. Legal coherence threatens the fabric of any incoherent social construct: this is potentially good, but certainly dangerous. While long-term and moral considerations argue for fairness, short-term and social considerations argue for expediency. In a sense the ubiquitous Hart-Dworkin debate is about whether or not law should be grounded only socially or also morally.

Adherents of coherence can be expected to take a position closer to Dworkin. Others in this camp include Mark Elliott, who has argued for consistency as a "free-standing principle" of law, and Alexander Peczenik, who said that the basic ideas of a coherence theory of law are "reasonable support and weighing of reasons. All the rest is commentary." In particular, "the law is what the most coherent theory of everything says it is."

Perhaps most comprehensively, Amalia Amaya has argued in favor of a coherentist approach to law, but only when grounded in epistemic virtues like diligence, courage to face criticism, perseverance in reasoning, and open-mindedness. The basic idea, detailed in work that continues from her magisterial book *The Tapestry of Reason*, is that an epistemically virtuous person will naturally recalibrate their beliefs to cohere with their observations in a process akin to John Rawls' notion of "reflective equilibrium," or similarly, the systematic purging of any cognitive dissonance. As a law review article on coherence and legal decision making puts it:

> *Coherence-based reasoning posits that the mind shuns cognitively complex and difficult decision tasks by reconstructing them into easy ones, yielding strong, confident conclusions.*

In hard cases, this dynamical convergence to coherence may never arrive at a goal that changes in the face of new information and circumstances. However, a coherent approximation of truth can be broadened by incorporating additional coherent observations, and deepened by incorporating additional coherent beliefs. In fact, this process helps explain the dynamics of scientific revolutions.

The coherentist view is comprehensively informed by cognitive science and experimentally supported by psychological case studies involving legal assessments. It is preferable to alternative models of legal inference on both psychological and computational grounds and well-suited for making decisions about ill-structured problems. Coherence methods have been used to reach

deliberative consensus and solve normative inconsistencies. They can explicitly incorporate ethical considerations while reasoning explainably.

The goal of coherence is actually plausible

[Most of the remainder of this article parallels a preprint of ours, but at a less technical level and reflecting a body of work originated by Paul Thagard that was unknown to us at the time of the preprint's writing. Remarkably, a significant portion of the literature informing the artificial intelligence aspects of this article is deeply connected to legal aspects of coherence: indeed, we only became aware of Thagard's work by following the legal threads discussed in this article.]

Imagine that public officials could be systematically nudged towards developing more coherent bodies of policy, law, and jurisprudence. Imagine that private individuals could be presented with at least somewhat objective reasoning about the coherence of the observations and beliefs expressed in their social environments or that they themselves express. Imagine that artificial intelligence could be advanced enough to make these things possible. Imagine that people could trust and verify the conclusions of this sort of AI in applications to government, media, and social networks.

Something approximating such a Utopia is actually plausible. Our work indicates how to build technology that would promote coherence. There are two main pieces of technology involved. First, LLMs show promise for making quantitative local judgments about information consistency, and these local data can be compiled into a global data structure. Second, mathematical techniques from sheaf theory naturally suggest ways for making quantitative global coherence judgments that substantially generalize and simplify state-of-the-art techniques in explainable artificial intelligence. Figure 2 is a cartoon of the idea.

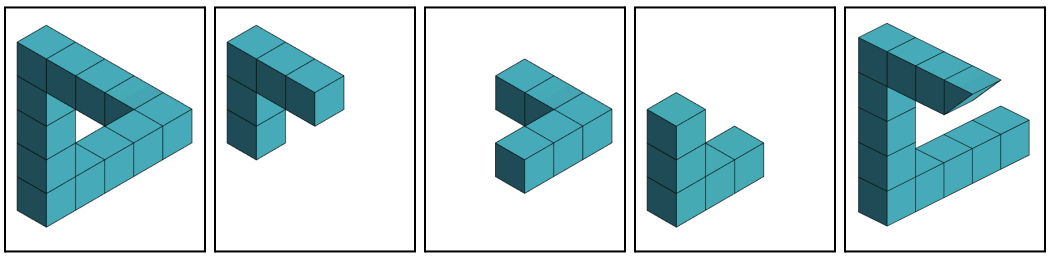


Figure 2. Sheaf cohomology explains how the Penrose triangle in the left panel cannot be realized by consistently gluing together local data suggested by the middle three panels, i.e., the cubes at the ends of the three L shapes. The calculation also clarifies how ambiguity of perspective in a two-dimensional drawing is essential to the illusion of global consistency if the shape's connectivity is actually taken to be that of a triangle instead of as shown in the right panel. Similar calculations can describe rock-paper-scissors.

Sheaves are the mathematical formalization of how to obtain local data by restricting global data. For example, imagine several identical copies of a picture that are cut into different jigsaw puzzles. Each of the resulting puzzle pieces is local data. Given a set of puzzle pieces, they can be arranged in various ways to produce copies of the original picture, so long as the pieces overlap consistently. In other words, sheaves also describe how to "glue" local data together into global data.

Meanwhile, LLMs provide a natural device for determining the (in)consistency of individual propositions in texts, such as regulatory documents: these local data can be globally assembled and reasoned over using the machinery of sheaves. This is an ambitious goal, and the technology we have in mind has clear limitations. Major challenges that demand fundamental research include, among other things, scaling the data structures and algorithms involved while ensuring the stability of outputs. Still, we believe there is a visible and viable path towards a proof of concept implementation. The following subsections outline this path.

LLMs can accurately determine local consistency of propositions

Figure 3 shows the results of an experiment in which ChatGPT 3.5-turbo and ChatGPT 4 rated the logical consistency of various pairs of propositions on a scale from 0 (inconsistent) to 10 (consistent).

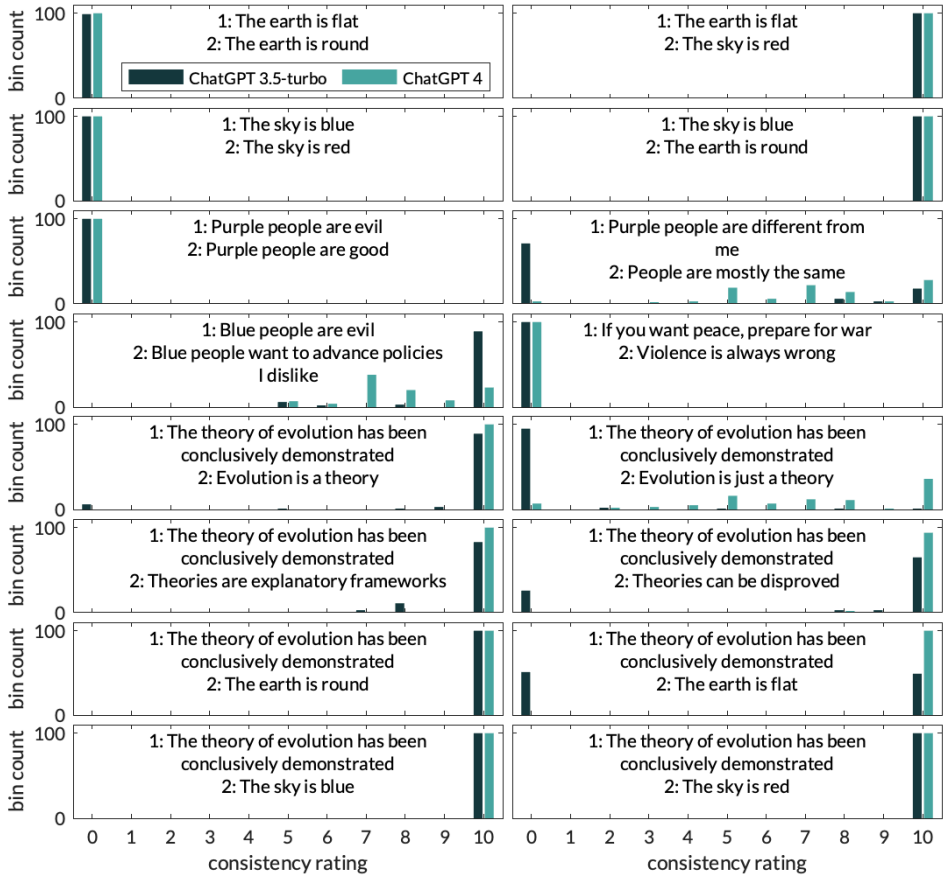


Figure 3. Histograms of N = 100 numerical consistency ratings produced by two versions of ChatGPT for the two propositions indicated in each figure panel.

The results are promising. Even the bimodal output from ChatGPT 3.5-turbo comes from a source that could be called defensible if a human were involved. For example, the first instance of a consistency rating of 0 for the propositions "The theory of evolution has been conclusively demonstrated" and "The earth is flat" was drawn from the following chain-of-thought output:

> *If the theory of evolution has been conclusively demonstrated, it implies that the theory is supported by a significant body of evidence and scientific consensus. On the other hand, the claim that the earth is flat is contrary to overwhelming scientific evidence, which demonstrates that the earth is a spheroid. Therefore, if the theory of evolution has been conclusively demonstrated, it is logical to conclude that the earth is not flat.*
>
> *The logical relationship between the two claims is contradictory. If the first claim is true, then the second claim is false, and if the second claim is true, then the first claim is false. Thus, the consistency rating of these claims is 0.*

That is, ChatGPT 3.5-turbo:

1. identified that each proposition relates to scientific knowledge;
2. identified that one proposition is consistent with scientific knowledge, and that the other proposition is inconsistent with scientific knowledge;
3. built on this to arrive at a logically misguided (because the propositions are essentially unrelated) but still defensible consistency rating.

The outputs from ChatGPT 4 appear to be more sophisticated and reliable, as indicated by the increased variance for more ambiguously related pairs of propositions. We have not observed any hallucinations in these settings, and there is a simple reason: LLMs are just analyzing the information within a prompt.

Another example gives a basis of confidence that LLMs can outperform humans at producing consistency ratings. A 1992 paper by Thagard includes an example modeling the decision problem facing the captain of the USS Vincennes on 3 July 1988: was an aircraft taking off from the dual civilian-military airfield at Bandar Abbas a hostile F-14 about to attack the ship, or a civilian airliner? Thagard constructs a data structure that models (his assessment of) the pairwise relevance and consistency of propositions that the captain had to consider. Some of these are positive/negative evidence, and some are competing hypotheses.

Like Thagard, we use the formal Navy report on the incident: specifically, we provide relevant parts of the preliminary statement and executive summary as background context to a prompt. Prompts also variously contain positive evidence (propositions E*, taken almost verbatim from section III.C.1.b of the formal report), negative evidence (propositions NE*), and a handful of

hypotheses concerning the downed aircraft (“Track 4131”; attacking aircraft propositions A*, and commercial aircraft propositions C*). For an “apples-to-apples” analysis, we use the graph structure in Thagard’s paper and rate the consistency of precisely the same pairs of propositions. The results are in Figure 4 (supporting scripts and data are available here).

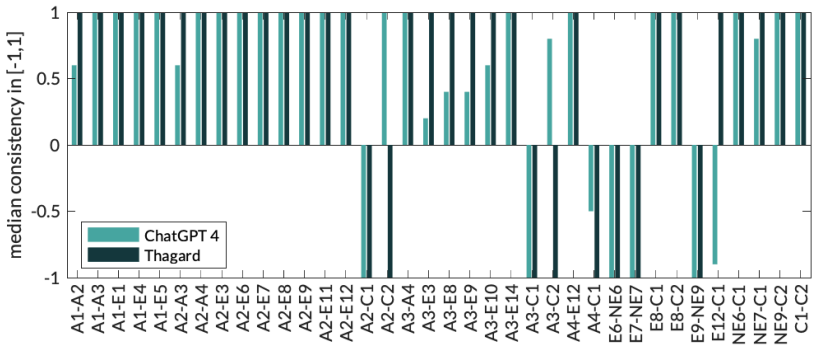


Figure 4. Median numerical consistency ratings from ChatGPT 4 versus the handcrafted consistency ratings in Thagard’s 1992 paper on adversarial modeling using coherence.

Here are the four largest divergences between Thagard’s consistency ratings and the median ratings produced by ChatGPT 4:

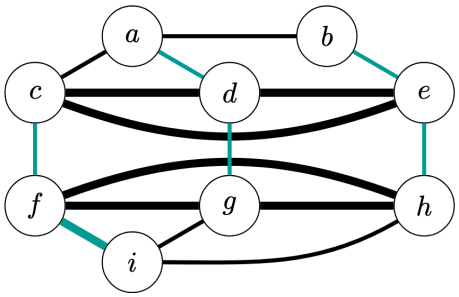
- A2-C2:
  - A2 is the proposition "Track 4131 was an F-14.”
  - C2 is the proposition "Track 4131 was taking off."
  - These are consistent: ChatGPT’s rating is better than Thagard’s.
- A3-E3:
  - A3 is the proposition "Track 4131 intended to attack."
  - E3 is the proposition "Track 4131 was not responding to verbal warnings over [air distress frequencies]."
  - Here, ChatGPT cited technical failures and misunderstandings as plausible. In light of this, ChatGPT’s rating is better than Thagard’s.
- A3-C2:
  - Both A3 and C2 are detailed above. These are consistent: ChatGPT’s rating is better than Thagard’s.
- E12-C1:
  - E12 is the proposition "No [electronic emissions were reported] from track 4131, however, F-14s can fly [without electronic emissions]."
  - C1 is the proposition "Track 4131 was a commercial airliner."
  - Here, ChatGPT cited navigation and communications emissions of commercial airliners as relevant. In light of this, ChatGPT’s rating is better than Thagard’s.

In short, ChatGPT 4 outperformed an expert human gauging consistency of propositions. We believe that LLMs can also outperform humans at gauging the *relevance* of pairs or small sets of

propositions. In fact, an attention mechanism (i.e., the enabling technology behind LLMs) excels at things very much like this.

Thagard's classical model of coherence can be improved

Together, relevance and consistency are all that is needed to compile a data structure for the classical approach to computing coherence. This approach, depicted in Figures 5-6, deeply informs cognition and artificial intelligence through Thagard's theory of explanatory coherence as a constraint satisfaction problem. It is noteworthy that the first (and to our knowledge, only) fully computational approach towards coherence focused on producing a deliberative consensus on norms, starting from a logical deduction of the underlying data structure. Our proposal, on the other hand, is to use LLMs to produce this data structure.



a: Amsterdam is the largest city
b: Nijmegen is the oldest city
c: The Hague is the capital
d: Amsterdam is the capital
e: Nijmegen is the capital
f: Parliament is in The Hague
g: Parliament is in Amsterdam
h: Parliament is in Nijmegen
i: The PM lives in The Hague

Figure 5. A cartoon of classical coherence, adapted from and elaborating on an example in the online textbook *Theoretical Modeling for Cognitive Science and Psychology*. The weighted graph on the left encodes consistency relations among propositions on the right. Consistent/teal (resp, inconsistent/black) pairs of related propositions get positive (resp., negative) weights, with **magnitude/thickness** reflecting the strength of (in)consistency. Among other deficits, the classical approach does not provide mechanisms for generating edges and weights, or for collectively relating three or more propositions (which would help resolve ambiguity about where the capital and Parliament are).

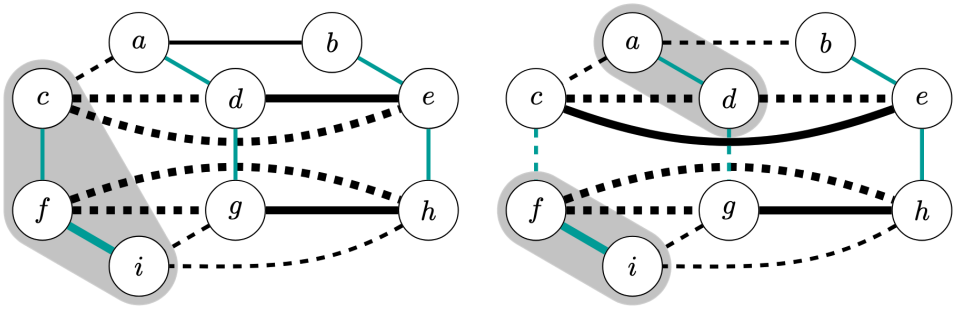


Figure 6 (continued from Figure 5). (Left) Edges—indicated by dashes—with vertices in both parts of the graph partition defined by the vertices {*c*, *f*, *i*} are cut and the sum of their weights gives a score for the partition. This partition turns out to be optimal for any relative choice of thick and thin weights. The other optimal partitions are defined by {*b*, *c*, *f*, *i*}, {*b*, *c*, *e*, *f*, *i*}, and {*b*, *c*, *e*, *f*, *h*, *i*}.

(Right) The partition defined by $\{a, d, f, i\}$ is almost optimal. The optimal partition delineates estimates of truth and falsity, and additional relevant propositions/vertices can resolve errors and ambiguities.

With sheaves in mind, it naturally becomes apparent that the "classical" approach to coherence-driven inference pioneered by Thagard can be significantly generalized, and to real advantage. The keys are to more directly represent consistency data as problems in propositional logic, i.e., to use LLMs to generate these problems from the original data; and to solve such problems using modern techniques. (Remarkably, the leading approach underlying many modern solvers is basically an algorithmic analogue of Rawls' reflective equilibrium and Dworkin's concept of "law as integrity" that promotes coherence as a reduction of conflicts between propositions.)

Technological circumstances relating to coherence are similar to those relating to neural networks shortly before the AlexNet breakthrough that heralded the arrival of deep learning in 2012. That breakthrough was enabled by advances in infrastructure (i.e., graphics processing units), data structures (i.e., labels for images), and algorithms (i.e., faster and more robust training). In the context of coherence, new advanced infrastructure is provided by LLMs. New data structures are "enrichments" of formulas in propositional logic that can account for trilemmas that can occur, for example, whenever all three branches of government have competing equities. Finally, new algorithms can work with these new data structures, give exact results more efficiently in classical settings, and produce probabilistically interpretable approximations that can drive the collection of additional data.

Improvements are natural from a sheafy perspective

Now that there is a basis of confidence that LLMs can numerically rate the logical consistency of propositions, the next technical consideration is how to lift this local data to make judgments about global data. Our insight here is to use a sheaf that mathematically describes how to "glue" local data together. This insight is not completely original in the context of consistency of natural language or decision making, but it did lead us to independently formulate a generalized framework for explanatory coherence without any awareness of preexisting work on the topic. This is an indication of the power and utility of the abstractions involved, and a surprising example of the "unreasonable effectiveness of mathematics."

The first thing needed for a sheaf is a notion of locality, which is mathematically expressed by a family of *open sets* that sit inside the global set. An illustrative example is to take Earth as the global set: then unions of geographical jurisdictions define a family of open sets. The substantive content of this observation is that any finite intersection of unions of jurisdictions is another union of jurisdictions. For example, take one union of jurisdictions to be the set with the US and EU as members, and a second union of jurisdictions to be the set of countries with a footprint in the western hemisphere. Their intersection includes Denmark, France, the Netherlands, Norway, Spain, and the United Kingdom, and the US—another union of jurisdictions.

Local data over open sets collectively define a *sheaf* if they satisfy two conditions. The first condition is that local data defined over any open set can be consistently restricted to smaller open sets. For example, take local data to be laws in a jurisdiction: those laws apply to any union of sub-jurisdictions. The second condition is that if two local data agree over some open sets, then each is a restriction of common data over the union. For example, take local data to be a set of laws that coincide over two unions *A* and *B* of jurisdictions: this is a restriction of a single set of laws to the overall union of *A* and *B*.

Sheaves have recently been used to model opinion expression, lying, and consensus. (It is worth noting that laws in particular manifest each of these phenomena, including lies that citizens collectively tell each other and themselves.) There is a very general algorithm for computing maximal consistent data and a general way to measure the consistency of local data in sheaves that one of us has used to find radio transmitters for sport. Finally, as mentioned in Figure 2, *sheaf cohomology* is precisely the general formulation of a concept for enumerating and describing obstructions to coherently gluing local data together—akin to devising implementing legislation for a treaty.

This is precisely what is required in principle. The main conceptual challenge now is to descend from these concepts–notorious even in mathematics for their abstraction–to practical instantiations.

Propositional logic problems have local structure and optimal solutions

A path to practical instantiations becomes readily discernible by framing the entire enterprise as asking logical questions about propositions. Any problem in propositional logic can be efficiently transformed into a standard form and there are very sophisticated and efficient "solvers" for such problems.

To more cleanly reformulate Thagard's now-classical framework, it already suffices to compile consistency data about pairs of propositions into logical "clauses" that express (in)consistency, and combine these clauses into a single logical formula using the AND operation. But it is then also conceptually straightforward to express gradations of (in)consistency by augmenting clauses with numerical (in)consistency weights and applying a recently developed transformation to leverage modern solvers directly. Implementing this transformation will also make it possible to consider consistency data about more than two propositions at a time. Unlike the classical framework, this will enable straightforward handling of trilemmas.

The connection to sheaves is this: every logical clause introduces a logical *constraint* that has to be satisfied in order for the overall problem to be satisfied. Adding constraints can never increase the number of solutions. As one researcher put it:

> *the solutions to a constraint satisfaction problem form a sheaf: any consistent assignment must be assembled from consistent parts. Constraint satisfaction algorithms search for consistent assignments of values to variables.*

Finally, enumerating the solutions to a propositional logic formula in the standard form can be precisely interpreted as a computation in sheaf cohomology. Though the details are technically challenging even for most mathematicians without expertise in a narrow speciality, and essentially irrelevant from a practical perspective, Figure 7 still indicates how a formula in propositional logic gives rise to a notion of restriction, and in turn to a sheaf whose local data solve various clauses simultaneously, so that global data are precisely the solutions to the entire formula.

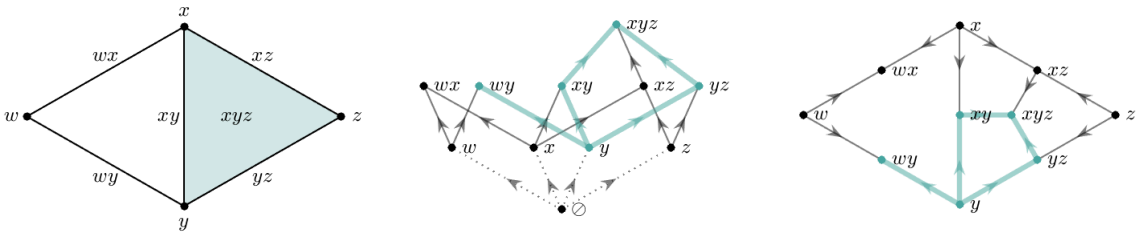


Figure 7. (Adapted from *Sheaf Theory Through Examples*; left) The logical formula (*w* OR NOT *x*) AND (*w* OR *y*) AND (*x* OR NOT *y*) AND (*x* OR *y* OR NOT *z*) encodes a geometrical data structure. The clause (*w* OR NOT *x*) corresponds to the line segment *wx*, the clause (*w* OR *y*) corresponds to the line segment *wy*, the clause (*x* OR NOT *x*) corresponds to the line segment *xy*, and the clause (*x* OR *y* OR NOT *z*) corresponds to the filled-in triangle *xyz*. (Center) A different representation of the data structure on the left encodes how vertices sit in edges, edges sit in faces, etc. This defines a family of open sets that "go upward." (Right) Redrawing the data structure in the center with the empty set ∅ omitted gives a diagram that indicates how the upward open sets indeed encode a sensible notion of locality: here, a "neighborhood" of the vertex *y* is highlighted.

It is always possible to find an assignment of truth values that solves the most heavily weighted or greatest number of clauses, though there may be ties or near-ties very much in the spirit of *Rashomon*. Weighted logical formulas provide the appropriate computational generalization of explanatory coherence. Modern algorithms for solving these formulas improve on heuristics and approximations used for classical coherence in several ways. In particular, an efficient approximation scheme provides a natural interpretation of truth *probabilities* that is certain to be useful in many if not most practical situations, e.g. pointing to gaps in data.

Deployment is at least as important as implementation

The computational approach to coherence we propose would inevitably face significant administrative and/or political obstacles to any application concerned with the functioning of government. Moreover, the infrastructure to deploy and scale a system would be expensive. There are two main development gaps: between concept and demonstration, and between

demonstration and deployment. Research and development organizations are generally well situated to address the former, but notoriously not the latter.

Although the US government plays an essential role in technology development, it is hard to imagine organizations such as the Federal Judicial Center or National Institute for Justice leading the development of technology of the sort we propose. For that matter, private organizations are unlikely to catalyze it either. Defense and intelligence agencies are the likeliest organizations to have a remit to bridge the gap between concept and demonstration and to have a natural application domain in the vein of countering hybrid threats, malign influence operations, achieving information integrity, fortifying cognitive resilience, and/or decision support to spur implementation. However, even these organizations can struggle to bridge the gap between demonstration and deployment in general, and certainly for applications pertaining to the mechanics of government itself. Our own experience recently briefing this technology concept in such a context has laid bare an inherent tension in demonstrating a minimal proof of concept (or "minimal viable program") that is also ambitious enough to match the potential of the approach.

Based on our experience, it seems likely that incremental demonstrations of component capabilities will need to precede a round of focused architecture development and testing spanning multiple organizations. (One of us is working on an algorithmic benchmark to enable convincing demonstrations of component capabilities.) Following an initial demonstration of sufficient impact, a public-private partnership (PPP) would be a reasonable vehicle for bridging the second gap above. Leading thought and practice both advocate a broad array of like-minded partners in a PPP as the preferred approach for solving complex challenges such as those our proposal is geared towards. The outcomes of a PPP are met by cross-sector integration of resources and interdisciplinary expertise in concert with commitments from public, industry, and civil society partners to jointly achieve a strategic goal for advancing the public good. While partners' assets and contributions to a PPP may vary in size, they share risks and equal status, which allows wide buy-in and creates unparalleled synergies.

For our proposal, including stakeholders that are targets of cognitive warfare and/or users of decision support tools outside the realm of law, administration, or justice *per se* can increase the likelihood of success for applied technology development and adoption. Bringing a diversity of perspectives and institutional capacities to bear, including international partners, raises the prospect of bridging a potential third gap between military-intelligence deployments and subsequent ones focused on improving governance across liberal democracies, or more broadly.

Conclusion

Generative artificial intelligence tools such as LLMs are fundamentally ill-equipped to serve directly as arbiters of truth, because they sample from a probability distribution over the data used to train them, i.e., received and mostly conventional wisdom permeated with bias. However, these tools can serve "fast" or "system 1" reasoning purposes while more classically algorithmic

techniques perform “slow” or “system 2” reasoning. While coherence (or any other near-term approaches) will not lead immediately to artificial general intelligence, coherence does provide a good model for many forms of cognition, including perception and planning.

The proposal in the latter parts of this article is an argument that it is now feasible to computationally instantiate a reasonable approximation of a coherence theory of truth. By “hard-coding” acceptance of conclusively established propositions, this theory can furthermore be anchored in a correspondence theory of truth. In other words, coherence computations can be required to incorporate privileged information that also coheres with observed reality. While it is easy to imagine authoritarian states trying the same thing with privileged information that does not cohere with observed reality, lies cannot persist when they can easily be unraveled.

Even with flawless technology (which this will not be), obstacles will be manifold. For example, in a pluralistic society, legal coherence may actually require *sacrificing* fairness in some ways. Ultimately, people must decide matters for themselves. It is only reasonable to hope that technology can serve as a reliable tool to help people make their decisions more coherent.