Argupedia

WHITEPAPER · VERSION 1.0

Der Global Debate Evaluation Standard

Ein Standard, um politische Argumente sichtbar, prüfbar und vergleichbar zu machen — und die Lücke zwischen demokratischer Legitimität und Entscheidungsqualität zu verkleinern.

VERSION 1.0
STAND Juni 2026
HERAUSGEBER Argupedia

AUTOR Julian Immanuel Ohm, M.Sc.

$$\text{Score} = V \times I \times P \div 10$$

ZUSAMMENFASSUNG

Demokratien gewinnen ihre Legitimität daraus, dass viele entscheiden. Doch viele zu beteiligen garantiert nicht, dass gut entschieden wird. Zwischen der *Legitimität* einer kollektiven Entscheidung und ihrer *Qualität* klafft eine Lücke, die nicht an bösem Willen liegt, sondern an der Art, wie Menschen Argumente verarbeiten.

Der Global Debate Evaluation Standard (GDES) ist ein Versuch, diese Lücke zu verkleinern. Er beruht auf einer einzigen strukturellen Annahme: Jede politische Entscheidung ist eine Wahl zwischen zwei Zukünften — einer mit und einer ohne die Maßnahme. Ein Argument ist dann nichts anderes als die Behauptung einer Differenz zwischen diesen Zukünften. GDES zerlegt diese Behauptung in drei unabhängig prüfbare Größen — **Value**, **Impact** und **Plausibility** — und macht die Bewertung, die ohnehin in jedem Kopf stattfindet, explizit, sichtbar und korrigierbar.

Dieses Whitepaper beschreibt das Problem, aus dem GDES entsteht, die Herleitung des Standards aus ersten Prinzipien, seine Eigenschaften und Grenzen sowie seine Anwendung in der strukturierten Politik-Enzyklopädie Argupedia.

01 — DAS PROBLEM

Urteilsbildung versagt nicht zufällig.

Wer ein politisches Argument bewertet, verarbeitet Hinweisreize — *Cues*. Ob das Urteil gut wird, hängt davon ab, welche Cues durchdringen und wie stark sie gewichtet werden.

Manche Cues tragen tatsächlich Information über die Sachfrage: Belege, logische Struktur, dokumentierte Wirkungen. Das sind **direkte Cues**. Andere sagen wenig über die Sache, aber viel über ihr Umfeld: das Selbstvertrauen des Sprechers, seine Parteizugehörigkeit, die emotionale Aufladung der Botschaft, der Konsens der eigenen Gruppe. Das sind **indirekte Cues**. Der grundlegende Befund der Urteilsforschung — von Brunswiks Linsenmodell bis zur modernen Sozialpsychologie — lautet: Wie stark ein Cue ein Urteil prägt, entspricht systematisch nicht dem, wie viel er tatsächlich zur richtigen Antwort beiträgt. Indirekte Cues sind oft zugänglicher und wirken stärker, als ihr Informationsgehalt rechtfertigt.

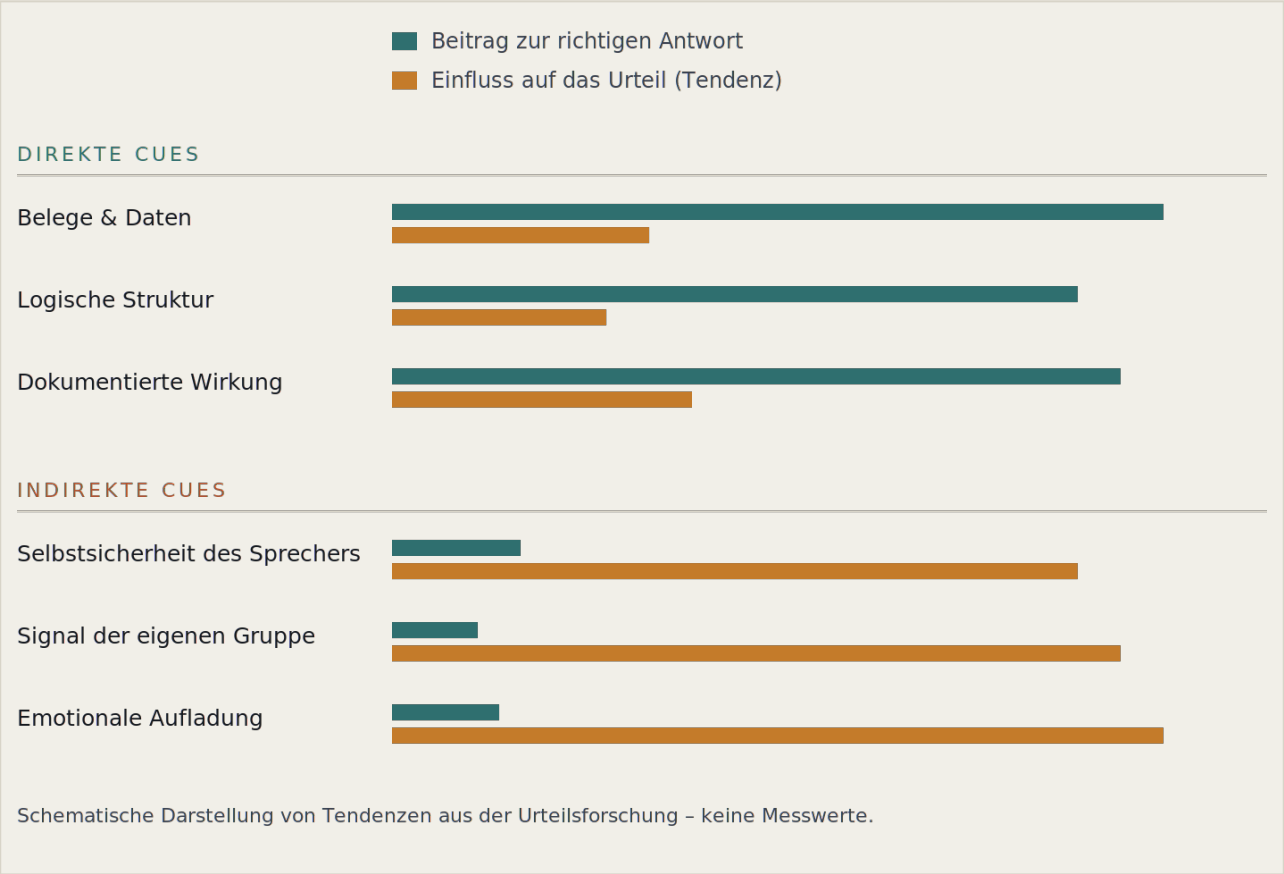

Abb. 1 — Direkte Cues tragen viel zur richtigen Antwort bei, beeinflussen das Urteil aber tendenziell wenig; bei indirekten Cues ist es umgekehrt. Schematische Darstellung von Tendenzen aus der Urteilsforschung, keine Messwerte.

Der Grund liegt in der Architektur unseres Denkens. Lange wurde sie als zwei getrennte Systeme beschrieben — schnelles, intuitives gegen langsames, prüfendes Denken. Neuere Arbeiten fassen sie genauer als **Kontinuum** von eher automatischer zu eher kontrollierter Verarbeitung. Drei Tendenzen genügen, um das Kernproblem zu erfassen.

INTUITIV

Schnell — und empfänglich für die falschen Signale

Am automatischen Ende urteilen wir mühelos und sofort. Dieses Denken reagiert tendenziell stark auf indirekte Cues und neigt dazu, aus dem gerade Verfügbaren eine kohärente Geschichte zu formen. Was fehlt, fällt selten auf.

PRÜFEND

Sparsam eingesetzt — und selten unparteiisch

Kontrollierte Verarbeitung könnte korrigieren, ist aber aufwendig und wird sparsam aktiviert. Springt sie an, arbeitet sie häufig eher als Anwalt denn als Richter: Sie sucht Gründe für die Seite, zu der wir ohnehin neigen.

SPEICHER

Eng — und leicht überfordert

Die Kapazität für bewusste Verarbeitung ist deutlich begrenzt. Unstrukturierte Information — und so kommt politische Debatte meist daher — überlastet sie leicht, bevor eine gründliche Prüfung überhaupt beginnen kann.

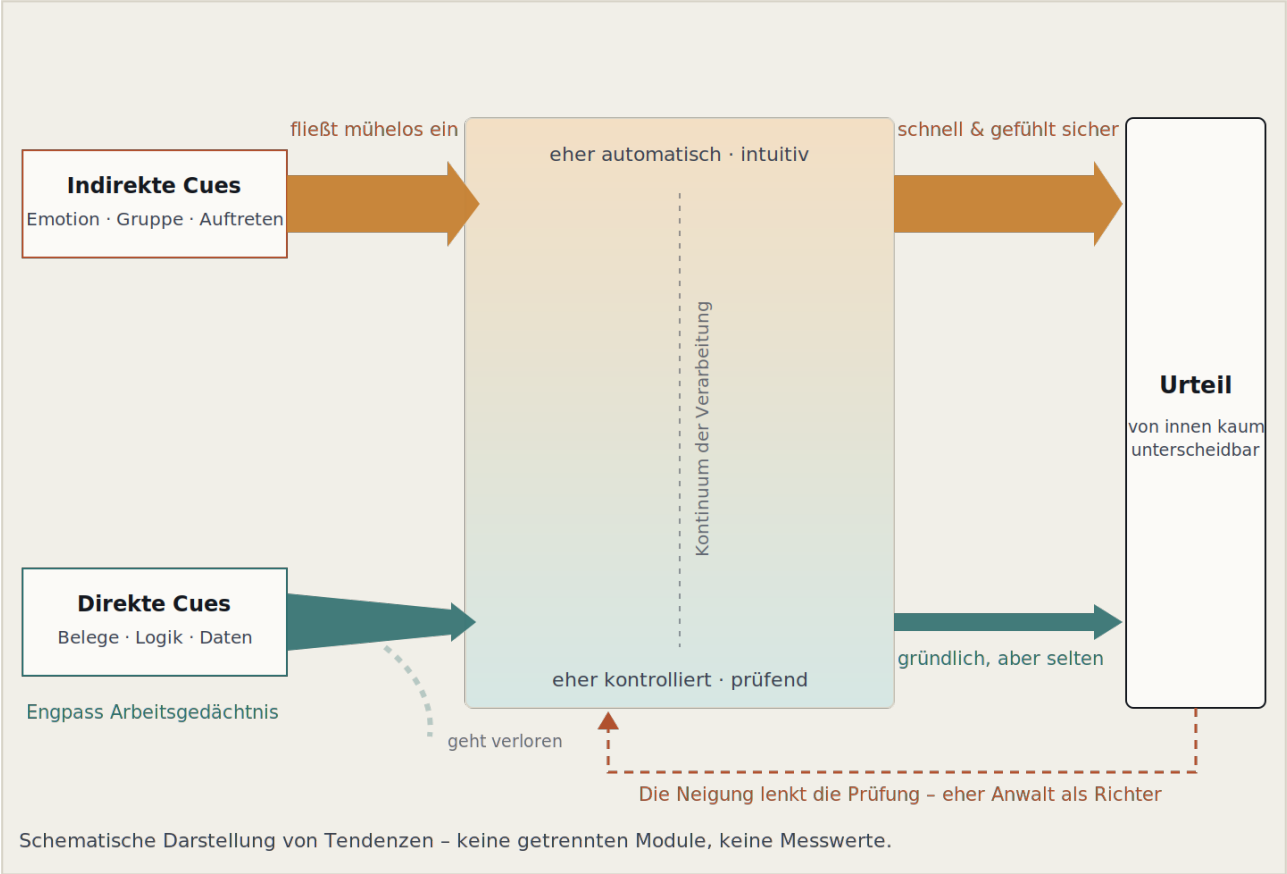

Abb. 2 — Indirekte Cues fließen mühelos in das automatische Ende und erzeugen ein schnelles, gefühlt sicheres Urteil. Direkte Cues müssen durch den Engpass des Arbeitsgedächtnisses; die gründliche Prüfung ist selten und wird zudem von der bestehenden Neigung gelenkt. Beide Wege münden in ein Urteil, das von innen kaum unterscheidbar ist.

Das ist eine Vereinfachung — die Forschung kennt weitere Mechanismen, und im Zusammenspiel von Gruppen verschärfen sich diese Tendenzen oft noch. Aber es ist der Kern. Seine heikelste Folge: Ein Urteil, das überwiegend von indirekten Cues getragen wird, kann sich von innen genau so anfühlen wie ein gut begründetes. Aus der ersten Person ist der Unterschied kaum zu erkennen.

> *Guter Wille allein schützt davor kaum. Die Bewertung von Wert, Wirkung und Glaubwürdigkeit findet bei jeder Entscheidung ohnehin statt — nur meist unsichtbar, unbewusst und damit schwer korrigierbar.*

---

02 — DIE GRUNDIDEE

Jede politische Entscheidung ist eine Wahl zwischen zwei Zukünften.

Aus dieser einen Beobachtung lässt sich der gesamte Standard ableiten — ohne weitere Vorannahmen über Ethik oder Weltbild.

Eine Maßnahme wird umgesetzt oder nicht. Damit stehen zwei Zukünfte zur Wahl: **F1**, die Welt, in der die Maßnahme greift, und **F0**, der realistische Verlauf, wenn sie ausbleibt. F0 ist dabei nicht der heutige Zustand, sondern die beste Schätzung, wie sich die Lage *ohne* die Maßnahme entwickeln würde — der Status quo trajektoriell weitergedacht.

Ein Argument, auf seinen logischen Kern reduziert, ist dann die Behauptung einer Differenz zwischen F1 und F0 entlang einer benannten Dimension — ein *Delta*. „Diese Reform senkt die Wartezeiten" heißt: In F1 sind die Wartezeiten kürzer als in F0. Genau dieses Delta, seine Größe und seine Glaubwürdigkeit, ist der Gegenstand, den GDES bestimmt, bewertet und prüft.

NOTATION

F1	Zukunft mit der Maßnahme
F0	Realistische Baseline ohne die Maßnahme (Status quo weitergedacht)
Delta (Δ)	Behaupteter Unterschied zwischen F1 und F0 entlang einer Dimension
V	Value — normatives Gewicht der Dimension (0–10)
I	Impact — normalisierte Größe des Deltas (0–10)
P	Plausibility — Eintrittswahrscheinlichkeit angesichts der Belege (0–10)

03 — DER STANDARD

Vom Delta zur Bewertung — vier Schritte.

Die Formel ist kein Ausgangspunkt, sondern ein Ergebnis. Sie folgt zwingend, sobald man ein Argument als Delta zwischen zwei Zukünften begreift.

SCHRITT 1

Impact — das Delta benennen

Jedes Argument behauptet einen Unterschied zwischen F1 und F0: weniger Emissionen, höhere Mieten, kürzere Wartezeiten. Der erste Schritt macht die Behauptung explizit und bestimmt ihre Größe — den Impact.

SCHRITT 2

Plausibility — den Erwartungswert bilden

Vorhersagen sind unsicher. Ein großes Delta mit schwacher Beleglage darf nicht so zählen wie ein kleineres mit starker. Der Impact wird daher mit seiner Eintrittswahrscheinlichkeit gewichtet. Was entsteht, ist im Kern ein Erwartungswert (Impact × Plausibility). Komplexere Debatten erlauben verfeinerte Formeln; das Prinzip bleibt dasselbe.

SCHRITT 3

Normalisieren — vergleichbar machen

Deltas kommen in verschiedenen Einheiten: Tonnen CO_2, Euro, Lebensjahre. Damit Argumente einer Debatte nebeneinander bestehen, wird der erwartete Impact gegen einen festen Kalibrierungsanker normalisiert — das Größte, was in diesem Feld auf dem Spiel steht.

SCHRITT 4

Value — mit unseren Werten gewichten

Nicht jede Dimension ist gleich wichtig. Der normalisierte Erwartungswert wird mit dem Value gewichtet: wie sehr uns die betroffene Dimension, gemessen an unseren Werten, am Herzen liegt. Hier — und nur hier — gehen Werte ein, offen und sichtbar statt versteckt.

Erst am Ende dieser Herleitung steht die Formel — nicht als Setzung, sondern als Konsequenz. Sie bündelt die vier Schritte zu einem einzigen Wert je Argument:

ARGUMENT-SCORE

$$\text{Score} = V \times I \times P \div 10$$

V, I, P jeweils 0–10 · Score-Bereich 0–100

Die Division durch 10 hält den Score auf einer handhabbaren Skala von 0 bis 100. Die Multiplikation ist keine Designwahl, sondern Entscheidungslogik: Impact × Plausibility ist der erwartete Effekt; mit Value gewichtet ergibt sich, wie stark ein Argument für oder gegen die Maßnahme wiegt.

Hinein- und herauszoomen

Die Formel ist nur ein Teil des Werkzeugs. Ihr eigentlicher Nutzen liegt darin, dass man die Granularität frei wählen kann. Jede Einzelbewertung — jedes V, I und P jedes Arguments — bleibt separat sichtbar, anfechtbar und korrigierbar. Wer einer Plausibility nicht traut, prüft genau sie. Zusammengerechnet ergeben die Scores aller Pro- und Contra-Argumente den **Debate Score**: einen schnellen Überblick über eine komplizierte Debatte, hinter dem die vollständige Begründung jederzeit aufklappbar bleibt. So entsteht beides — ein Gesamturteil und volle Nachvollziehbarkeit bis zur einzelnen Annahme.

---

04 — WAS GDES IST — UND WAS NICHT

Ein Interface für eine Bewertung, die ohnehin geschieht.

GDES verlangt keinen neuen Denkprozess. Wer eine Politik unterstützt oder ablehnt, hat Wert, Wirkung und Glaubwürdigkeit bereits bewertet — implizit, mit all den Abkürzungen, die die Forschung beschreibt. Der Standard macht diese ohnehin laufende Bewertung explizit und legt sie dorthin, wo sie geprüft, angefochten und verbessert werden kann.

Er ist dabei ein **Diskurswerkzeug, kein Entscheidungswerkzeug**. GDES verändert, was Entscheidende sehen — nicht, wer entscheidet. Die demokratische Entscheidung bleibt, wo sie hingehört. Und er verspricht keine Perfektion: Die ideale Entscheidung — die ein perfekter Beobachter mit vollständiger Information träfe — bleibt Richtpunkt, nicht Anspruch. GDES verschiebt reale, fehlbare Urteilsbildung messbar in ihre Richtung.

Auf eine bestimmte Moralphilosophie legt der Standard niemanden fest. Er macht nur die minimale Annahme, dass Politikentscheidungen Entscheidungen zwischen Zukünften sind — vereinbar mit religiösen, deontologischen, tugendethischen wie utilitaristischen Positionen. **Value** ist der Ort, an dem diese Unterschiede offen verhandelt werden, statt versteckt zu wirken. Und wo die Beleglage zu dünn ist, erzwingt GDES kein Urteil: *Aussetzen* ist ein legitimes Ergebnis, kein Versagen.

---

05 — ANWENDUNG IN ARGUPEDIA

Vom Standard zur Enzyklopädie.

Argupedia wendet GDES systematisch auf Politikbereiche an, sodass nicht nur einzelne Debatten bewertet, sondern Vorschläge innerhalb eines Feldes vergleichbar werden.

Jeder Politikbereich (Bereich) erhält eine gemeinsame Ausgangslage: eine einzige, sorgfältig begründete **F0-Baseline** und einen festen **Kalibrierungsanker**. Beide gelten für alle Vorschläge im Feld. Dadurch werden Bewertun-

gen über Vorschläge hinweg vergleichbar — eine Eigenschaft, die eine debattenweise, jeweils neu kalibrierte Auswertung nicht leisten kann.

Auf dieser Grundlage wird jeder Vorschlag **symmetrisch** rekonstruiert: Die Pro-Argumente ergeben sich aus der Sache selbst, die Contra-Argumente werden als die stärkste Einwandsfigur eines informierten Kritikers rekonstruiert — nicht als Strohmann. Jedes Argument trägt seine Herkunft (Provenienz) und seine Belege mit Qualitätsstufen. Die Bewertung selbst wird **append-only** versioniert: Frühere Einschätzungen werden nicht überschrieben, sondern als Versionsstand erhalten, sodass die Entwicklung eines Urteils nachvollziehbar bleibt.

DIE AUSWERTUNG EINES BEREICHS IN KÜRZE

1 · Baseline	Eine gemeinsame F0-Trajektorie für das gesamte Feld festlegen.
2 · Anker	Den größten Einsatz im Feld als fixen Kalibrierungsanker wählen.
3 · Vorschläge	Jeden Vorschlag als Delta gegen dieselbe Baseline fassen.
4 · Symmetrie	Pro- und Contra-Seite mit Provenienz und Belegen rekonstruieren.
5 · Score	V, I, P je Argument gegen den festen Anker — Argument- und Debate-Score.
6 · Version	Bewertung append-only versionieren; nichts überschreiben.

06 — WARUM JETZT

Im Loop bleiben.

Die Luftfahrt hat eine teure Lektion gelernt: Automation erzeugt eine eigene Gefahr. Wer lange nur überwacht statt selbst zu fliegen, verliert mit der Zeit genau die Fähigkeiten, die im entscheidenden Moment gebraucht werden — das „Out-of-the-loop"-Problem. Die Antwort war nie, den Menschen zu ersetzen, sondern das System so zu bauen, dass menschliches Urteilen verlässlich funktioniert.

Künstliche Intelligenz tut dem demokratischen Diskurs Vergleichbares wie der Autopilot dem Fliegen: Sie erzeugt flüssige, selbstbewusste Argumente in einem Maßstab, den kein Mensch mehr prüfen kann, und lädt dazu ein, das eigene Urteilen abzugeben. Die Verlockung, der Maschine zu vertrauen, ist dieselbe — und dieselbe Gefahr droht: dass die Fähigkeit zur Prüfung verkümmert, gerade wenn sie am nötigsten ist.

GDES überträgt die Lehre der Luftfahrt auf den Diskurs. Es zerlegt Behauptungen in Dimensionen, die Bürgerinnen und Bürger selbst bewerten können, macht Begründungen sichtbar und prüfbar — und erzwingt genau die aktive Auseinandersetzung, die zu umgehen die Technologie verführt. KI wird damit nicht zum Ersatz des Urteils, sondern zum Instrument, das menschliches Urteilen im Loop hält.

07 — GRENZEN UND OFFENE FRAGEN

Was der Standard nicht leistet.

Ein Standard, der seine eigenen Grenzen verschweigt, verstößt gegen das Prinzip, dem er dient. Die wichtigsten Vorbehalte offen benannt:

- **Bewertungen bleiben Urteile.** V, I und P sind begründete Schätzungen, keine Messungen. Der Standard macht sie nur explizit und anfechtbar — er macht sie nicht objektiv. Zwei sorgfältige Menschen können bei gleicher Beleglage unterschiedlich scoren; GDES verschiebt den Streit lediglich an die richtige, sichtbare Stelle.
- **Die Baseline F0 ist selbst eine Vorhersage.** Welche Zukunft ohne die Maßnahme einträte, lässt sich nie sicher wissen. Eine falsch gesetzte Baseline verzerrt jedes Delta im Feld. Argupedia begegnet dem mit einer gemeinsamen, offengelegten F0 — beseitigt das Problem damit aber nicht, sondern macht es nur verhandelbar.
- **Quantifizierung kann Scheingenauigkeit suggerieren.** Eine Zahl wirkt präziser, als sie ist. Der Zoom-Mechanismus — jede Zahl bis zu ihrer Begründung aufklappbar — soll dem entgegenwirken, hebt das Risiko aber nicht vollständig auf.
- **Nicht alles Politische zerfällt sauber in Deltas.** Fragen von Identität, Würde oder Verfahrensgerechtigkeit lassen sich nicht immer als Differenz zwischen zwei Zukünften fassen. Dort liefert GDES bestenfalls einen Teilbeitrag — und sollte seine Grenze eingestehen, statt sie zu kaschieren.
- **Der Standard ist nur so gut wie seine Anwendung.** Symmetrische Rekonstruktion, faire Anker und ehrliche Plausibilities verlangen Disziplin. Schlecht angewandt kann GDES bestehende Verzerrungen mit einer Aura von Objektivität versehen — das gefährlichste Versagen, gegen das nur Transparenz und Anfechtbarkeit schützen.

---

08 — WISSENSCHAFTLICHER HINTERGRUND

Das Framework erfindet nichts.

Es fügt etablierte Befunde aus Jahrzehnten der Urteils-, Kognitions- und Demokratieforschung zusammen. Neu ist die Architektur, nicht die Bausteine.

Brunswik & Hammond `Lens Model · Social Judgment Theory`	Der grundlegende Befund: Wie stark ein Cue ein Urteil prägt, entspricht systematisch nicht dem, wie viel er tatsächlich zur richtigen Antwort beiträgt.
Kahneman u. a. `Dual-Process-Forschung · WYSIATI`	Intuitives Urteilen formt aus dem gerade Verfügbaren eine kohärente Geschichte und übergeht, was fehlt. Neuere Arbeiten fassen die Zwei-Systeme-Metapher als Kontinuum von automatischer zu kontrollierter Verarbeitung.
Petty & Cacioppo `Elaboration Likelihood Model`	Zwei Verarbeitungsrouten — zentral (Inhalt, Belege) und peripher (Sprecher, Gefühl, Oberfläche). Selbst inhaltlich starke Argumente können peripher verarbeitet werden, ohne Prüfung ihrer Triftigkeit.
Zaller `Receive–Accept–Sample-Modell`	Politische Botschaften müssen erst empfangen und verarbeitet werden, bevor sie wirken — die Trennung von Empfang und Annahme als zwei getrennte Hürden.
Sweller `Cognitive Load Theory`	Begrenztes Arbeitsgedächtnis: Überlastung blockiert die Verarbeitung valider, aber komplexer Information, bevor sie das Urteil erreicht.
Lodge & Taber · Kahan	

`Motivated Reasoning · identitätsschützende Kognition`	Bestehende Überzeugungen und Gruppenidentität lenken die Cue-Gewichtung: Bestätigendes wird verstärkt, Widersprechendes abgewertet — besonders bei politisch aufgeladenen Fragen.
Wilson & Brekke `Mental Contamination`	Urteile können von unerwünschten, nicht introspektierbaren Einflüssen geprägt sein — und fühlen sich von innen wie gut begründete an.
Asch · Janis `Konformität · Gruppendynamik`	In Gruppen verschärft sich das Problem: Konsenssignale werden zu dominanten indirekten Cues, abweichende valide Information wird unterdrückt.
Tetlock `Good Judgment Project`	Vorhersagequalität ist messbar und trainierbar: Strukturierte Verfahren, Kalibrierung und Feedback verbessern Urteile dauerhaft — der empirische Beleg, dass die Lücke schließbar ist.
Fishkin `Deliberative Polling`	Beraten Bürgerinnen und Bürger unter strukturierten Bedingungen mit ausgewogener Information, verändern und verbessern sich ihre Urteile messbar.

Über dieses Dokument. Dieses Whitepaper beschreibt GDES in der Fassung V1.0. Der Standard ist bewusst als lebendes Dokument angelegt: Definitionen, Pipeline und Skalen können sich in künftigen Versionen präzisieren. Rückmeldungen, Einwände und Gegenbeispiele sind ausdrücklich willkommen — sie sind genau die Cues, auf die ein Standard zur Bewertung von Argumenten selbst hören sollte.