

行政文書に、見えない印が付くことが決まった

——生成AIの電子透かしと政府AI基盤「源内」。残された時間を、日本は測れない

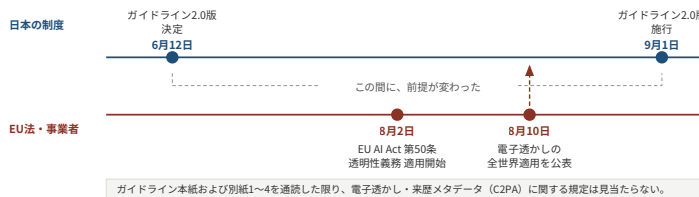
一般財団法人 日本危機管理研究所 | 2026年8月 | 読了目安 3分 | 記述は2026年8月14日時点の公表情報に基づく

全府省庁18万人が対象の政府AI基盤「源内」にはAnthropic社のClaudeが搭載され、令和8年度中には国会答弁の草案を生成するAIの開発が予定されている。同社は、生成テキストに機械可読な印を全世界で付与すると公表した。猶予はあるが、その終期を決めるのは日本ではない。本紙は詳細版レポートの要点を2ページに要約したものです。記述はすべて公表された一次資料に基づき、出典を併記しています。

1. なぜ今か —— 3つの日付と、1つの判断機会

図1 制度の決定と、技術環境の変化

日本のガイドラインが決定された後、施行される前に、前提となる技術環境が変わった。



① 8月2日 EUの透明性義務が発効した。

EU AI Act第50条の適用が開始され、生成AIの提供者には、AI生成・加工コンテンツに機械可読なマークを付す義務が生じた。違反企業には最大1500万ユーロまたは全世界年間売上高の3%の制裁金が科される可能性がある。

〔出典〕ITmedia NEWS（2026年8月4日）

② 8月10日 事業者が、全世界への適用を公表した。

米Anthropic社は、AI「Claude」が生成するテキストに機械可読な電子透かしを埋め込む方針を公表した。同社はEU AI Act第50条2項の行動規範に署名しており、適用範囲はEU域内に限られず、日本を含む世界全体に及ぶ。印はテキストそのものに織り込まれ、コピー＆ペーストしても移動し、一部の編集を経ても残る。ただし対象は2026年8月2日以降にリリースされるモデルであり、それ以前のモデルには経過措置がある。

〔出典〕Anthropic「How Claude marks AI-generated content」（2026年8月10日更新）

③ 9月1日 日本のガイドラインが施行される。

デジタル庁「行政の進化と革新のための生成AIの調達・利活用に係るガイドライン」（DS-920）2.0版が施行される。決定は6月12日であり、上記②の公表はその後である。本紙および別紙1～4を通読した限り、同ガイドラインに電子透かし・来歴メタデータ（C2PA）に関する規定は見当たらない。

〔出典〕デジタル庁 デジタル社会推進会議幹事会決定（2026年6月12日）／デジタル庁「ガバメントAI『源内』」

要点。これはガイドラインの不備ではない。6月12日の決定時点で、8月10日の公表は存在しなかった。問題は、施行を目前にして前提が変化したにもかかわらず、対応の手続きが見えないことにある。

ただし、日本が決められることが一つある。

そして8月末。デジタル庁の資料は、源内について「令和9年度以降に本格的利用（各省庁が予算措置）」と明記している。この8月の概算要求は、18万人規模の基盤を各省が自らの恒常経費として計上するか否かを定める最初の回である。上の3つは日本に対して起きたことであり、これは日本が決めることである。

〔出典〕デジタル庁「（参考資料）ガバメントAI源内の展開状況」2026年5月

2. 日本の空白 —— 2つ、そして1つの問い

空白① 公文書管理制度が、この印を想定していない。

職員が生成AIで作成した文書は、組織的に用いるものとして保有されれば行政文書となる（内閣府大臣官房公文書管理課長通知・令和7年2月14日）。しかし現行の整理が求めるのは「正確性を確保するため必要な確認を経ること」であり、AI利用の事実そのものを記録・開示することではない。制度が記録を求めている事実が、技術によって自動的に記録される状態が生じている。

これは将来の想定ではない。政府AI基盤「源内」は全府省庁約18万人を対象とした実証段階にあり、搭載モデルとしてClaude Sonnet 4.6・4.5、Claude Haiku 4.5とNova Liteが公表されている。「公用文を校正」アプリ（文化庁ガイドライン準拠）は実装済みであり、校正に用いただけでも印は付く。さらに「国会答弁作成支援AI」——国会質問に対する答弁の草案を生成するアプリ——が令和8年度中の開発予定として公表されている（令和7年度補正予算44.0億円の内訳に計上、受託：NTTドコモビジネス、2026年4月～2027年3月）。答弁作成を支援するAIスキルセットは2026年3月から提供開始済みである。Anthropic社は、AWS等のクラウド経由でも透かしが適用されるとしており、源内が構築されるガバメントクラウド経由は対象範囲に含まれる。

そして開示請求への対応業務そのものが、AI適用先として名指しされている。同資料は「文書・記録管理（文書登録、保管・整理、情報公開）」の課題として「情報公開請求で対象文書の特定に膨大な時間」を挙げ、「国会対応・議会対応（質問通告対応、答弁案起草）」も高度な生成AIが求められる業務として位置づけている。

想定される問い：開示請求された行政文書に印が含まれる事実は、開示・説明の対象となるか。検出手段を持たない官庁は、その事実を認識できない。認識できない事項について、説明責任は成立するか。

猶予の終期を、日本側は決められない。公表資料が搭載モデルを挙げるのは「2026年3月時点では」という留保付きであり、本紙執筆時点の構成を外部から検証する手段はない。その3月時点の構成であればClaude 3モデルはいずれも8月2日より前のリリースであり、印はまだ付いていないと推定される。しかし猶予が終わる経路は二つある——(1) 源内のモデル更新（デジタル庁は「今後、モデルの追加・更新を予定」と公表）、(2) 事業者による既存モデルへの遡及適用（同社は経過措置の下で対応追加作業中と公表）。いずれか早い方で猶予は終わり、その時期はどちらも事業者側の判断と作業進捗に依存する。終期のみならず、現在地さえ外部からは測れない。透かしの適用時期と、答弁草案生成AIの実装時期は、同じ令和8年度中に重なる。

空白② その印を、日本政府が独立に検証できない。

検出の仕組みは今後公開される技術文書で示されるとされ、現時点で第三者による検証手段は提供されていない。日本の行政文書の来歴に関する情報を、判定できるのは生成元の企業のみという状態にある。

検討に値する問い 適用除外とした領域の取扱い

DS-920ガイドライン2.2.1は、特定秘密・重要経済安保情報・安全保障等の機微情報を扱う政府情報システムを、ガイドラインの全部の適用対象外としている。適用除外は日本の制度上の判断であり、事業者のマーキング方針を左右しない。**ただし源内は「機密性2情報まで入力可能」とされ、適用除外領域そのものを扱う設計ではない。**したがってこの問いが残るのは源内ではなく、**各府省庁が個別に調達する生成AIシステム**である。DS-920の適用除外は、まさにその個別調達に及ぶ判断だからである。実態を把握し、確認の要否を判断できるのは政府自身である。

本紙が主張していないこと。テキスト透かしはモデルの単語選択をわずかに偏らせる方式であり、文書の内容を外部に送信するものではない。論点は情報漏えいではなく、日本の制度が要求していない記録が、日本が読めない形式で行政文書に残ることである。

3. 何を求めるか ― 3点

提言① 機械可読マーキングの位置づけを、施行までに整理する。

手段	先進的AI活用アドバイザーボードによるガイドラインの見直し（法改正を要さない）
所管	デジタル庁
時期	2026年9月1日の施行にあわせて

整理すべき論点は3つ。(1) 生成物に付与される機械可読な印を、行政文書の属性としてどう扱うか。(2) 印の存在が、情報公開請求への対応において意味を持つか。(3) 調達仕様書において、マーキングの有無および検出手手段の提供を要求事項とするか。

提言② 検出手手段の独立確保を、調達要件に位置づける。

この種の透かしは秘密鍵による検出を前提とする。求めるべきものは次の3点に分解できる。**(1) 政府への検出手手段（API等）の提供／(2) 検出鍵の第三者預託**（AISはIPA内に設置。2025年12月の総理指示でAI担当大臣・経産大臣に抜本的強化・200人体制が指示済み。所管：内閣府・経済産業省／調達要件はデジタル庁）／**(3) 調達契約への検出協力義務の明記。**

これは特定事業者への要求ではない。Google社は既にGeminiの本番環境でテキスト透かしを運用しており、同社の実装でも一般利用者が検出スコアを得る手段はないと指摘されている。「事業者のみが検出できる」構造は社固有ではなく、この技術に共通する。したがって全事業者に対する共通の調達基準として設計すべきである。

提言③ AI利用の記録に関する考え方を、萎縮ではなく整理として示す。

この印は、AI利用を告発する道具としては機能しない。校正や翻訳に用いただけでも付き、大幅に書き直せば消える（詳細版 図2）。この非対称を放置すれば、検出手手段の普及後、印の有無をめぐる非難の応酬が生じ、結果として職員がAI利用そのものを避けるようになる。ガイドラインが掲げる「利活用促進とリスク管理を表裏一体で進める」方針にとって、最も避けるべき帰結である。**必要なのはAI利用を抑制する規律ではなく、どこまでの利用がどのように記録され、その記録が何を意味しないのかを、あらかじめ明示する整理である。**

4. 国会で問うなら、この4問

質問主意書は、内閣が原則7日以内に答弁する義務を負う（国会法第75条）。9月1日施行・8月末概算要求という日程に対し、最も速い確認手段となりうる。いずれも政策の当否ではなく事実関係の確認であり、答弁は可能である。

問1	源内が搭載する生成AIモデルのうち、提供事業者が機械可読なマーキングの対象として公表しているものはあるか。対象外である場合、(a) 今後のモデル追加・更新に伴い対象となる見込み、(b) 事業者による既存モデルへの遡及適用の見込みについて、政府はそれぞれの時期を把握しているか。把握していない場合、照会する方針はあるか。
問2	政府は、行政文書に含まれる機械可読なマーキングを独立に検出する手段を保有しているか。保有していない場合、取得する方針はあるか。所管はいずれの機関となるか。
問3	開示請求の対象となった行政文書に当該マーキングが含まれる場合、その事実は開示・説明の対象となるか。検出手手段を保有しない場合、当該事実を認識できないこととなるが、行政機関情報公開法上の説明責務との関係でどのように整理されるか。
問4	令和8年度中に開発が予定されている国会答弁作成支援AIについて、生成される答弁の草案に機械可読なマーキングが付与される可能性はあるか。政府はこの点を開発の要件として検討しているか。当該草案を基に作成された答弁書におけるマーキングの取扱いについて整理はなされているか。

5. 想定される問いと、答え

Q. 現行モデルは対象外なら、まだ関係ない話では。

A. 公表情報の範囲ではそう推定されます。ただし搭載構成が公表されているのは2026年3月時点までで、現在の状態を外部から確認する手段はありません。だから今が期限です。猶予が終わる経路は二つあります――源内のモデル更新と、事業者による既存モデルへの遡及適用です。**後者は日本政府に決定権がなく、時期を事前に知る手段も公表されていません。**源内がモデルを一つも入れ替えなくても猶予は終わります。確定している日程は9月1日の施行だけです。

Q. 文書の中身が外部に送られるということか。

A. 違います。透かしはモデルの単語選択をわずかに偏らせる統計的な信号であり、通信を伴いません。情報漏えいの問題ではなく、日本の制度が要求していない記録が、日本が読めない形式で残るという問題です。

Q. EUの法律であって、日本には関係ないのでは。

A. 事業者が全世界に適用すると公表しています。日本の行政職員が業務で利用した場合も、その出力に印は付きます。日本の法令が要求していない印が、EU法に由来する事業者の自主的約束に基づいて付与される、という構造です。

Q. 印が付けば、AIが書いたことが分かるのでは。

A. 分かりません。事業者自身が限界を明示しています。人間が書いた原稿を校正・翻訳・要約させただけでも印は付き、逆に大幅に編集・言い換えされた文章や、短い文章では検出できない場合があります。検出が示すのは「処理に関与した可能性」までです。

Q. 職員のAI利用を制限すべきという話か。

A. 逆です。整理がないまま検出手手段が普及することこそが、萎縮を招きます。印が付くことは不名誉ではない――この理解が先に共有されていれば、利活用は妨げられません。本提言は利活用を促進するための整理を求めるものです。

Q. 民間の自主的取組に委ねればよいのでは。

A. 委ねた結果が現状です。検出手手段を持つのは生成元の企業のみであり、政府が独立に検証する手立てがありません。行政文書の来歴という、本来国が管理すべき情報についての話です。

詳細版・関連レポート 本紙は四層AI主権フレームワーク系列レポートの第3稿にあたる。

詳細版：「行政文書に、見えない印が付くことが決まった――EU AI Act第50条の域外適用と、9月1日施行ガイドラインの空白」（2026年8月）

関連：「四層フレームの破綻事例と対策」（2026年7月30日）／「『実世界ネイティブAI』という賭け」（2026年7月22日）

発行：一般財団法人 日本危機管理研究所 執筆：船山 美保 / © 2026 本資料は政策検討・引用目的での転載・配布を自由に行っていただけます（出典明記のお願い）。