

Legal Dilemmas of AI-Extensible Products

Analytical Case Studies on Multi-Doctrinal Attribution and Cross-Jurisdictional Liability

Wanhong HUANG

huangwanhong.g.official@gmail.com

Working Draft v0.2 — May 11, 2026

Abstract

AI-extensible products—software systems that allow end users to extend or customize functionality through natural-language prompts processed by embedded large language models—constitute a novel product paradigm at the convergence of platform design, user-programmability, and generative artificial intelligence. This paradigm creates a set of legal dilemmas that traditional legal doctrine, developed for human-to-human conduct mediated by deterministic tools, is not equipped to resolve. This article identifies and analyzes three structural sources of dilemma common across legal doctrines: *black-box causation* (the structural unknowability of causal contribution through opaque AI processing); *mixed culpability* (the coexistence of intentional, negligent, and innocent contributions among multiple users); and *cross-agent liability allocation* (the unsettled distribution of responsibility among product providers, users, and the AI system itself). Through four analytical case studies spanning single-user and multi-user, mono-jurisdictional and cross-jurisdictional scenarios, the article maps these dilemmas across seven legal doctrines—criminal law, tort law, copyright and intellectual property, contract law, data protection, administrative regulation, and private international law—and undertakes a comparative analysis across six jurisdictions: Germany, Japan, China, the United States, the United Kingdom, and the European Union. The article does not propose a unified doctrinal solution. Rather, it argues that the dilemmas are *structural* rather than incidental, that they affect multiple doctrines in homologous ways, and that any adequate regulatory response will require multi-doctrinal coordination, ex-ante regulation as the principal governance modality, and criminal law’s continued retreat to the role of *ultima ratio*. The article concludes with preliminary recommendations for legislators, product designers, and the legal scholarly community on how to engage these dilemmas without prematurely collapsing them into either over-regulation or under-regulation.

Keywords: AI-extensible products; black-box causation; mixed culpability; multi-user attribution; comparative criminal law; tort liability; AI regulation; private international law.

26 Executive Summary

27 Background and the Object of Study

28 A new product paradigm has emerged in the wake of widely deployed large language models: *AI-*
29 *extensible products*. These are commercial software systems—ranging from productivity suites and
30 family-management applications to development tools and decision-support platforms—that allow
31 end users to extend, customize, or repurpose the product’s functionality by issuing natural-language
32 prompts to an embedded AI component. Where previous generations of user-extensibility (spread-
33 sheet macros, scripting environments, visual workflow tools) operated through deterministic inter-
34 preters, AI-extensible products route user intent through a stochastic, opaque generative model. This
35 shift creates legal questions that existing doctrine was not designed to answer.

36 The Three Structural Dilemmas

37 This article identifies three structural sources of legal dilemma that recur across multiple doctrines
38 and jurisdictions:

- 39 1. **Black-box causation.** Causal attribution in law presupposes that the chain from conduct to conse-
40 quence is, at least in principle, knowable. The architecture of contemporary AI systems makes this
41 assumption structurally untenable: not merely difficult or temporarily obscure, but *constitutively*
42 *unknowable* at the level of precise contribution. Doctrines that rely on causal proof—criminal
43 causation, tort causation, copyright contribution—are destabilized at their epistemological founda-
44 tions.
- 45 2. **Mixed culpability of multiple users.** When multiple users independently submit prompts that
46 collectively shape an AI output, their mental states are typically heterogeneous: some intentional,
47 some negligent, some entirely innocent. Traditional doctrines of complicity, joint tortfeasance, and
48 co-authorship presuppose either a common subjective state (intentional concert) or fundamentally
49 similar contributions. The mixed-culpability scenario sits outside both assumptions.
- 50 3. **Cross-agent liability allocation.** Three classes of agent are implicated in every AI-extensible
51 product harm: the product provider (Company B), the user or users, and the AI system itself.
52 Existing doctrine offers fragmentary tools—intermediary safe harbors, contributory liability, the
53 Winny doctrine of neutral tool providers, Section 230 immunity, the EU AI Act’s risk-graded
54 obligations—but no coherent framework that integrates these across doctrines or jurisdictions.

55 Methodological Approach

56 The article is structured as a multi-doctrinal, comparative-law analysis built on four analytical case
57 studies. The case studies are *constructed*, not drawn from actual judgments: each is designed to isolate
58 a specific configuration of intent, multiplicity, and jurisdictional scope, in the manner of German
59 *Fallkonstellation* or American legal hypotheticals. The seven doctrinal fields surveyed are criminal
60 law, tort law, copyright and IP, contract law, data protection, administrative regulation, and private
61 international law. The six jurisdictions analyzed are Germany, Japan, China, the United States, the

62 United Kingdom, and the European Union considered as a supranational regulatory regime.

63 Principal Findings

64 **The dilemmas are structural, not incidental.** The three sources of dilemma identified above mani-
65 fest in homologous form across doctrines: the problem of opaque causation appears in criminal cau-
66 sation *and* in tort causation *and* in the contribution test for joint authorship. This homology suggests
67 that no single-doctrine fix will succeed; the dilemmas are inherent in the product paradigm itself.

68 **Jurisdictions cluster into three regulatory paradigms.** Comparative analysis reveals three fam-
69 ilies of response: (i) *continental doctrinal parsimony* (Germany, Japan), which preserves doctrinal
70 coherence at the cost of leaving substantial gray zones; (ii) *constitutional and statutory protection of*
71 *intermediaries* (United States, especially under Section 230 and the First Amendment), which min-
72 imizes provider liability but offers little to victims; (iii) *ex-ante regulatory governance* (European
73 Union, United Kingdom), which front-loads compliance obligations and reduces reliance on case-
74 by-case adjudication.

75 **Intentional actors are reliably attributable across all jurisdictions.** Every jurisdiction studied
76 attributes responsibility to users who knowingly direct AI systems toward unlawful outputs, typi-
77 cally through the doctrine of indirect perpetration or its functional equivalents. The divergences arise
78 downstream of intent: in the treatment of negligent users, of provider liability, and of evidentiary
79 standards under black-box conditions.

80 **Evidence law may be more constraining than substantive law.** Substantive doctrines often
81 *permit* attribution in difficult cases, but the evidentiary burden—establishing causation to a probabil-
82 ity sufficient for criminal conviction or civil liability—cannot be met under structural unknowability.
83 The article argues that this is not a defect to be remedied through evidentiary inversion, but a feature
84 of legal humility that substantive law should respect.

85 Preliminary Policy Recommendations

86 The article offers four interrelated recommendations, developed in detail in Chapter 8:

87 1. **Treat criminal law as ultima ratio.** Resist the temptation to expand criminal causation doctrines
88 or evidentiary presumptions to accommodate AI-mediated harms. Reserve criminal sanction for
89 clear intentional misuse; route negligent and ambiguous cases through civil and regulatory chan-
90 nels.

91 2. **Adopt risk-graded ex-ante regulation as the principal governance modality.** The EU AI Act's
92 tiered approach—differentiating obligations by use-case risk level—offers a template that can be
93 adapted to other jurisdictions without requiring doctrinal upheaval. This approach is compatible
94 with both common-law and civil-law systems.

95 3. **Establish safe-harbor compliance frameworks for product providers.** Provide clear, achiev-
96 able standards under which providers who implement specified safety measures (content filtering,
97 abuse detection, transparent ToS, incident response) are shielded from secondary liability, akin to
98 but more structured than current DMCA-style safe harbors.

99 4. **Sandboxed extensibility as a design norm.** Encourage product architectures that limit AI-mediated
100 extension to domain-specific languages or pre-defined operation sets, rather than open-ended code
101 generation. This protects users, providers, and third parties simultaneously by constraining the
102 space of possible harms ex ante.

103 **Scope and Limitations**

104 This article does not propose a unified doctrinal solution to the dilemmas it identifies. The dilemmas
105 may not admit of such a solution; legal systems operating across multiple doctrines and jurisdictions
106 may have to live with structural indeterminacy in this area, just as they have learned to live with sim-
107 ilar indeterminacies in mass-tort causation, transboundary environmental harm, and other domains
108 where scientific certainty outruns legal proof. The article's contribution is diagnostic and cartographic
109 rather than prescriptive: to map the terrain in a way that subsequent scholarship, adjudication, and
110 legislation can build upon. The case studies are constructed and analytical, not empirical; the juris-
111 dictional sample, while broad, is necessarily selective; and the technological reference point—the
112 current generation of large language models—will evolve, possibly rendering some specific claims
113 obsolete even as the structural dilemmas persist.

114 _____

115 Contents

116	Executive Summary	2
117	1 Introduction	9
118	1.1 A Modest Hypothetical	9
119	1.2 The Research Problem	9
120	1.3 Three Concepts and the Argument's Architecture	10
121	1.4 Methodology	11
122	1.5 Roadmap	11
123	2 AI-Extensible Products as a New Legal Object	12
124	2.1 The Genealogy of User-Extensibility	12
125	2.2 Defining the Paradigm	13
126	2.3 Distinction from Adjacent Concepts	13
127	2.4 Why This Paradigm Requires Legal Reconceptualization	14
128	3 Analytical Case Studies and the Dilemma Matrix	15
129	3.1 Methodological Preliminaries	15
130	3.2 Scenario α : Single-User Intentional Misuse	16
131	3.3 Scenario β : Single-User Negligent or Innocent Use	17
132	3.4 Scenario γ : Multi-User with Mixed Culpability	18
133	3.5 Scenario δ : Cross-Jurisdictional Multi-User Configuration	20
134	3.6 Scenario ε : A Six-Actor Multinational Configuration with Heterogeneous Culpability	21
135	3.6.1 Facts	21
136	3.6.2 The Procedural Pipeline	22
137	3.6.3 Actor-by-Actor Analysis	24
138	3.6.4 The Enforcement Reality	25
139	3.6.5 What Scenario ε Reveals	26
140	3.7 The Dilemma Matrix	27
141	3.8 Reading the Matrix: Three Patterns	28
142	4 Common Structural Dilemmas: Three Cross-Doctrinal Sources	29
143	4.1 Introductory Framing	29
144	4.2 Black-Box Causation	30

145	4.2.1	Traditional Causation and Its Epistemological Premise	30
146	4.2.2	Scientific Uncertainty as the Existing Exception	30
147	4.2.3	Why AI Black-Box Causation Is Different	31
148	4.2.4	Cross-Doctrinal Implications	32
149	4.3	Mixed Culpability of Multiple Agents	33
150	4.3.1	The Binary Structure of Traditional Complicity	33
151	4.3.2	Three Inadequate Existing Tools	34
152	4.3.3	Proposing a Layered Attribution Approach	35
153	4.4	Cross-Agent Liability Allocation	36
154	4.4.1	Three Classes of Agent, Three Tools, No Framework	36
155	4.4.2	The Insufficiency of Current Tools	37
156	4.4.3	Connecting to the Case Studies	38
157	4.5	The Homology Across Doctrines: A Synthesis	39
158	5	Doctrinal Dilemmas in Specific Legal Fields	39
159	5.1	Criminal Law	39
160	5.1.1	The Indirect Perpetrator Framework and Its Limits	39
161	5.1.2	Mixed Culpability in Penal Code Doctrine	40
162	5.1.3	The Causation Burden Under Black-Box Conditions	41
163	5.2	Tort Liability	42
164	5.2.1	Negligence, Foreseeability, and Joint Tortfeasance	42
165	5.2.2	The Joint Tortfeasance Problem in Scenario Gamma	43
166	5.2.3	Causation in Civil Litigation	44
167	5.3	Intellectual Property: Copyright and Beyond	45
168	5.3.1	The Authorship Question	45
169	5.3.2	Three-Layer Copyright in AI-Extensible Outputs	46
170	5.3.3	Beyond Copyright: Trade Secrets, Patents, Trademarks	47
171	5.4	Contract Law and ToS Architecture	47
172	5.5	Data Protection and Personal Information	48
173	5.6	Administrative Regulation	49
174	5.7	Private International Law	49
175	5.8	Other Doctrinal Implications	50
176	6	Comparative Analysis Across Six Jurisdictions	51
177	6.1	Introductory Framing	51

178	6.2	Germany	51
179	6.2.1	Governance Paradigm: Doctrinal Precision and Regulatory Embrace	51
180	6.2.2	Criminal Law: BGH Mittäterschaft Jurisprudence	52
181	6.2.3	Tort: BGB §823 and Product Liability	52
182	6.2.4	Administrative: AI Act Implementation	53
183	6.2.5	IP: §69 Series and AI Training	54
184	6.3	Japan	54
185	6.3.1	Governance Paradigm: Doctrinal Restraint and Soft-Law Preference	54
186	6.3.2	Criminal Law: The Winny Doctrine	54
187	6.3.3	Tort: Civil Code Art. 709 and Network Infringement	55
188	6.3.4	Administrative: AI 事業者ガイドライン	56
189	6.3.5	IP: Copyright Act Art. 30-4	56
190	6.4	China	57
191	6.4.1	Governance Paradigm: Codified Doctrine, Active Judicial Interpretation, Policy Steering	57
192	6.4.2	Criminal Law: Aiding-Information-Crime Offense	58
193	6.4.3	Tort: Civil Code Art. 1194-1197	58
194	6.4.4	Administrative: Generative AI Interim Measures	59
195	6.4.5	IP: Beijing Internet Court AI Image Case	60
197	6.5	United States	60
198	6.5.1	Governance Paradigm: Constitutional Protection of Intermediaries	60
199	6.5.2	Criminal Law: MPC §2.02 and Rosemond	61
200	6.5.3	Tort: Section 230 and Gonzalez	62
201	6.5.4	Administrative: NIST AI RMF and Sectoral Regulation	62
202	6.5.5	IP: Copyright Office Guidance and Thaler	63
203	6.6	United Kingdom	64
204	6.6.1	Governance Paradigm: Common Law Plus Active Statutory Reform	64
205	6.6.2	Criminal Law: Jogee and Common Law Complicity	64
206	6.6.3	Tort: Common Law Negligence and OSA	65
207	6.6.4	Administrative: OSA 2023 and AI White Paper	66
208	6.6.5	IP: UK Copyright Act §9(3) on Computer-Generated Works	66
209	6.7	European Union	67
210	6.7.1	Governance Paradigm: Comprehensive Ex Ante Regulation	67
211	6.7.2	Criminal Law: Member State Competence	67

212	6.7.3	Tort: Product Liability Directive 2024 and AI Liability Directive	68
213	6.7.4	Administrative: AI Act as the Centerpiece	69
214	6.7.5	IP: DSM Directive Art. 4 TDM Exception	70
215	6.8	Synthesizing the Comparison	70
216	7	Judicial Views and Cross-Jurisdictional Issue Analysis	71
217	7.1	Introductory Framing	71
218	7.2	Issue 1: Indirect Perpetration Through AI Tools	71
219	7.3	Issue 2: Foreseeability Standards for Negligent Users	72
220	7.4	Issue 3: Provider Liability and Safe Harbors	73
221	7.5	Issue 4: Black-Box Causation in Adjudication	74
222	7.6	Issue 5: Cross-Border Enforcement Realities	75
223	7.7	Three Regulatory Paradigms: A Synthesis	76
224	8	Policy Implications and Product Design Recommendations	77
225	8.1	The Humility of Diagnosis	77
226	8.2	Legislative Recommendations	78
227	8.2.1	Maintain Criminal Law as Ultima Ratio	78
228	8.2.2	Adopt Risk-Graded Ex-Ante Regulation	79
229	8.2.3	Establish Provider Safe-Harbor Frameworks	80
230	8.3	Product Design Recommendations	81
231	8.3.1	Sandboxed Extensibility as a Design Norm	81
232	8.3.2	Multi-Layer ToS Architecture	81
233	8.3.3	Audit Logging Within Privacy Constraints	82
234	8.3.4	Prompt-Level Isolation	82
235	8.4	International Coordination Recommendations	83
236	9	Conclusion and Open Questions	84
237	9.1	What the Article Has Argued	84
238	9.2	Methodological Reflection	85
239	9.3	Open Questions for Subsequent Scholarship	85
240	9.3.1	Doctrinal Refinement Questions	85
241	9.3.2	Comparative-Law Extension Questions	86
242	9.3.3	Technological Evolution Questions	86
243	9.4	A Closing Note on Humility	87

244 1. Introduction

245 1.1 A Modest Hypothetical

246 Consider a software product—call it *Hearth*—designed to help families manage their everyday
 247 lives: meal planning, household scheduling, shared shopping lists, gentle reminders about birthdays
 248 and medical appointments. The product is built on an architecture now common in consumer soft-
 249 ware: a stable core of curated functionality, surrounded by a user-facing layer that accepts natural-
 250 language requests. “Adjust this week’s grocery list to remove anything my mother is allergic to,” a
 251 user might say, and the product, mediated by an embedded large language model, will produce a re-
 252 vised list. “Draft a polite reply to my mother-in-law’s message about the holidays,” another user might
 253 request, and a draft appears. The product is, in the vocabulary developed below, an *AI-extensible*
 254 *product*: one whose functional surface is genuinely extended—not merely configured—through user
 255 prompts processed by a generative model.

256 Now imagine that four users of this product, unknown to one another, send prompts on the same
 257 afternoon. The first user, let us call her eA, intends to use the product’s reply-drafting capability
 258 to manipulate an elderly relative into transferring funds; she is fully aware of the unlawfulness of
 259 her plan. The second user, eB, is a software developer experimenting with prompts that probe the
 260 system’s behavior; she inadvertently elicits a response template that, in combination with other inputs,
 261 will turn out to be useful for the first user’s scheme. The third user, eC, asks a completely innocent
 262 question about persuasive writing for a high-school essay; her prompt happens to refine the model’s
 263 internal representations in a way that materially shapes the reply eA ultimately receives. The fourth
 264 user, eD, is a member of the same fraud ring as eA, working independently to refine prompt templates
 265 for the same campaign. The product’s underlying model, processing these inputs over a short window
 266 in which retrieval-augmented context and recent fine-tuning gradients interact, produces an output
 267 *E*: a reply, sent by eA to her relative, that succeeds in defrauding the victim of a substantial sum.

268 Who is responsible, and for what? The criminal lawyer reaches first for the doctrine of indirect
 269 perpetration, then for the law of complicity, then for the theory of cumulative causation. The tort
 270 lawyer asks about joint tortfeasance, about proximate cause, about the foreseeability of inadvertent
 271 contributions. The intellectual property lawyer wonders who holds rights in the reply draft. The
 272 data protection officer notes that eC’s innocent question may have exposed personal information
 273 of a third party in ways neither she nor the product provider can fully trace. The administrative
 274 regulator inquires whether the product provider has satisfied the obligations of the applicable AI risk-
 275 management framework. The conflict-of-laws specialist observes that the four users may be in four
 276 different jurisdictions, the provider in a fifth, the victim in a sixth, and the servers in a seventh.

277 None of these lawyers will produce a confident answer. Each will encounter, in the vocabulary
 278 of her own doctrine, a set of difficulties that this article calls *dilemmas*.

279 1.2 The Research Problem

280 This article asks three interlocking questions:

281 1. What is distinctive, as a matter of legal analysis, about AI-extensible products as a product paradigm,

and why does the existing architecture of liability doctrine struggle with this paradigm?

2. How do the principal legal doctrines—criminal law, tort law, intellectual property, contract law, data protection, administrative regulation, and private international law—each engage with the paradigm, and to what extent do their difficulties share a common structural source?

3. How do the major jurisdictions that confront these problems—Germany, Japan, China, the United States, the United Kingdom, and the European Union as a regulatory regime—differ in their responses, and what do these differences reveal about the possible shape of an international consensus or, in its absence, a stable pluralism?

The article does not aspire to a unified solution. The thesis advanced here is that the dilemmas it identifies are *structural*: they arise from features of the product paradigm itself, manifest in homologous form across doctrines, and admit of management rather than resolution. The article's contribution is, accordingly, diagnostic rather than prescriptive—to map the terrain so that subsequent legislation, adjudication, and scholarship can build upon a shared picture of where the difficulties actually lie.

1.3 Three Concepts and the Argument's Architecture

The article introduces three concepts that, taken together, articulate the structural sources of the dilemmas it analyzes.

Black-box causation refers to the structural unknowability of the causal contribution of any particular input to the output of a contemporary generative AI system. This is distinct from, and more radical than, the familiar problem of scientifically unascertained causation that has preoccupied courts in mass-tort and environmental contexts. Where scientific causation is in principle ascertainable given sufficient investigation, black-box causation is constitutively opaque: the mechanisms of attention-based language models do not yield to the kind of clean attribution that legal doctrine has traditionally presupposed.

Mixed culpability refers to the configuration in which multiple agents contribute to a single harmful outcome through inputs that reflect heterogeneous mental states—some intentional, some negligent, some innocent. Existing doctrines of complicity and joint tortfeasance presuppose either intentional concert or fundamentally similar contributions; the mixed-culpability scenario, which AI-extensible products make routine rather than exceptional, sits outside both assumptions.

Cross-agent liability allocation refers to the problem of distributing responsibility among the product provider, the user or users, and the AI system itself. Doctrine offers fragmentary tools—intermediary safe harbors, the Winny doctrine of neutral tool provision, Section 230 immunity, the EU AI Act's risk-graded obligations—but no coherent framework that integrates these across doctrines or jurisdictions.

These three concepts function as analytic lenses, not as a tidy taxonomy. Many of the dilemmas the article analyzes involve more than one of them simultaneously; some involve all three. The argument proceeds by working through analytical case studies, exposing the doctrinal difficulties each case poses, and showing how the three concepts illuminate the deeper structural sources of

those difficulties.

1.4 Methodology

The article combines three methodological registers. *Doctrinal analysis* predominates, in the continental and common-law traditions of close engagement with statutory text, judicial decisions, and scholarly commentary. *Comparative law* provides the horizontal axis, surveying six jurisdictions in parallel rather than treating one as default. *Case-study methodology*, in the analytical rather than the empirical mode, supplies the vertical axis: four constructed scenarios that recur throughout the article as test cases for each doctrine and jurisdiction.

The case studies are deliberately constructed—in the manner of German *Fallkonstellation*, American legal hypotheticals, or Japanese 設例—rather than drawn from actual judgments. This is a methodological choice with three justifications. First, no jurisdiction has yet produced an extensive body of case law on the specific configurations the article analyzes; the law is, in the most precise sense, still in formation. Second, constructed cases allow the analytical isolation of specific variables (multiplicity, intent, jurisdictional spread) that real cases rarely present in clean form. Third, the article’s contribution is intended to be of use to legal systems before the harms it describes become widespread; engaging the difficulties prospectively is more useful than waiting to react to actual victims.

1.5 Roadmap

The article is organized into nine chapters across four parts.

Part I (Chapters 1–3) introduces the problem, the product paradigm, and the four analytical case studies. Chapter 2 defines AI-extensible products and situates them in the genealogy of user-extensible software. Chapter 3 presents the four case studies and the dilemma matrix that organizes the remainder of the analysis.

Part II (Chapters 4–5) develops the doctrinal analysis. Chapter 4 treats the three structural sources of dilemma—black-box causation, mixed culpability, and cross-agent liability allocation—in their cross-doctrinal generality. Chapter 5 examines each of the seven doctrines in turn, treating criminal law, tort law, and intellectual property in depth, and the remaining four (contract, data protection, administrative regulation, private international law) more briefly.

Part III (Chapters 6–7) undertakes the comparative analysis. Chapter 6 surveys the six jurisdictions along the doctrinal axes established in Part II. Chapter 7 reorganizes the same material around specific judicial decisions and points of doctrinal divergence, building toward a taxonomy of regulatory paradigms.

Part IV (Chapters 8–9) draws preliminary policy implications and identifies open questions. Chapter 8 proposes design and regulatory principles consistent with the diagnostic analysis; Chapter 9 concludes by identifying lines of inquiry that this article opens but cannot close.

A reader pressed for time may consult the Executive Summary above, which provides a structured overview of the article’s principal claims and recommendations.

2. AI-Extensible Products as a New Legal Object

2.1 The Genealogy of User-Extensibility

The history of consumer software is, among other things, a history of the gradual displacement of the boundary between those who write software and those who use it. AI-extensible products are the most recent and most radical step in this displacement, but they are not the first; their distinctive features come into focus when they are situated in a longer trajectory.

The earliest commercially significant moment in that trajectory is the spreadsheet macro, exemplified by Lotus 1-2-3 in the early 1980s and generalized through Microsoft Excel's adoption of Visual Basic for Applications (VBA) in the 1990s. The macro converted the spreadsheet from a calculation tool into a small programming environment usable by accountants, analysts, and clerks who would never have considered themselves programmers. The legal reception of this development was muted: courts and legislators treated VBA scripts as either user data (where errors lay with the user) or as software components (where errors lay with the user-author), and the fact that the script ran inside Microsoft's product attracted no special liability theory.

A second wave arrived with extensible plugin architectures: Adobe Photoshop's filter API, Eclipse's plugin system, the browser extension ecosystems of Firefox and Chrome. Here a third party—neither the platform vendor nor the end user—contributed code that ran inside the host application. The legal regime that crystallized around this configuration was anchored in intermediary safe harbors (Section 230 of the US Communications Decency Act, the EU E-Commerce Directive of 2000, Japan's プロバイダ責任制限法 of 2001) and in contributory and vicarious infringement doctrines worked out in cases such as *A&M Records v. Napster* and *MGM v. Grokster*.

A third wave—low-code and no-code platforms such as IFTTT, Zapier, Airtable, and Notion—moved further outward, allowing users with no programming background at all to compose multi-step workflows from graphical primitives. Doctrinally, this wave attracted little distinct attention: the configuration space remained finite and predetermined, and courts continued to treat user-built workflows as ordinary user data.

The current wave—AI-extensible products built around large language models—differs in kind, not merely in degree. The earlier waves moved the locus of authorship outward while preserving the determinacy of execution: a VBA macro, a Photoshop filter, a Zapier workflow each behaves as a deterministic function of its inputs and the platform's defined operations. AI-extensible products preserve the outward displacement of authorship but abandon determinacy. The same prompt, issued twice, may produce materially different outputs; outputs may be shaped by recent inputs from other users via shared retrieval contexts; and the relationship between prompt and output is not, even in principle, analytically reducible to a sequence of human-readable operations. It is this combination—authorship by ordinary users, mediated by a non-deterministic and non-traceable process—that constitutes the new paradigm.

2.2 Defining the Paradigm

We adopt the following working definition. *An AI-extensible product is a software system in which a substantial portion of end-user-facing functionality emerges from the interaction between user-supplied natural-language prompts and an embedded generative AI model, rather than being pre-determined by the product's authors.* The definition is intentionally narrow on the dimension of *emergence*—the user is not merely choosing among options but supplying language that the system translates into novel product behavior—and intentionally inclusive on the dimension of *embodiment*, covering products in which the AI is implemented locally, called via API to a foundation model provider, or hybridized.

Three conditions must obtain simultaneously. First, there is a *stable core* of functionality that the provider designs, deploys, and warrants: a database, a document editor, a calendar, a code repository. Without a stable core there is only the underlying chat interface, which is a generative model rather than an extensible product. Second, the product accepts *natural-language prompts* from end users, not merely configuration parameters or graphical operations. The unbounded nature of natural language is essential: a configuration UI with a thousand options is not, on this definition, AI-extensible, because the option space is finite and pre-enumerated. Third, the prompts are processed by a *generative AI model whose outputs alter product behavior*, not merely produce auxiliary content. A product that uses an AI to suggest text the user will then paste into a separate workflow is not extended by the AI; a product whose database queries, scheduling decisions, or output drafts are generated by AI in response to user prompts is.

This definition makes three classes of product easy to classify and a fourth class genuinely difficult. Squarely *included* are systems such as Notion AI's database operations, GitHub Copilot when it directly modifies a repository, Claude Artifacts producing executable code, and the hypothetical Hearth of Chapter 1. Squarely *excluded* are pure chatbot interfaces (which lack a stable non-conversational core), traditional rule-based decision-support systems (which lack the natural-language extensibility), and translation or summarization utilities whose outputs are auxiliary text rather than behavior. *Borderline* are tools whose AI component is positioned just outside the product surface—a writing assistant that drafts replies but requires a user click to send, for example. Whether such tools fall within the paradigm depends on whether the AI's contribution is integral to the product's distinctive functionality or merely an upstream convenience. The borderline cases will not be resolved by definitional fiat; they will be resolved, in the end, by litigation and by the ordinary process by which legal categories absorb new technological forms. The legal interest of the present article does not turn on borderline classification but on the central cases, which are now abundant and will only proliferate.

2.3 Distinction from Adjacent Concepts

The category of AI-extensible products is most usefully approached by distinguishing it from four adjacent categories with which it is sometimes conflated.

General-purpose platforms (the Apple App Store, social media networks) host third-party content and applications, but ordinary users do not modify the platform itself. The legal regime that

has grown up around such platforms—intermediary safe harbors, content-moderation duties, marketplace liability—addresses a structurally different problem: how to allocate responsibility between the platform and content contributors who are typically distinct legal entities.

User-generated content platforms (YouTube, Wikipedia, Reddit) allow users to contribute content that other users consume. The output is content, not extended platform functionality. The user who posts a video on YouTube has not changed what YouTube does for any other user; she has added an item to a shared content corpus. AI-extensible products are distinct in that the user’s action changes *what the product does*, typically for that user and (in shared-context configurations) for subsequent users as well.

Configurable SaaS systems (Salesforce administration, Notion templates, Airtable formula systems) allow users to compose product behavior, but from a predetermined and finite set of operations. The configuration space is, in principle, enumerable; the system’s behavior remains analytically traceable as a function of explicit configuration choices. AI-extensible products break this finiteness: the prompt space is the space of natural language, which is unbounded, and the prompt-to-behavior function is opaque rather than rule-based.

Standalone AI generation tools (a chatbot used to draft text the user then pastes elsewhere; a code-generation tool used to scaffold a project the user then builds in a separate IDE) involve AI but do not extend a product. The user is the integrator, combining AI-produced artifacts with separately maintained workflows. Liability questions in this configuration tend to follow the path of the artifact: the AI provider for the artifact’s defects, the user for what she does with it.

These distinctions can be summarized along two axes: whether the user modifies the product itself or only its content, and whether the modification is bounded by predefined options or is open-ended through natural language. Table 1 sketches the resulting taxonomy.

	Bounded modification	Open-ended modification
Modifies content only	UGC platforms (YouTube comments, Reddit posts)	Standalone generative AI used outside any host product
Modifies the product itself	Configurable SaaS, plugin marketplaces, low-code tools	AI-extensible products (Notion AI, Copilot, Hearth)

Table 1: Adjacent paradigms distinguished from AI-extensible products.

The defining feature of AI-extensible products lies in the lower-right cell: the extension is open-ended, mediated by natural language and a generative model, and lives *inside and through* the product, changing what the product does for the user from the point of the extending prompt forward. It is this combination that gives rise to the legal dilemmas analyzed in the chapters that follow.

2.4 Why This Paradigm Requires Legal Reconceptualization

Each preceding generation of user-extensibility could, with effort, be accommodated by existing legal doctrine. A buggy spreadsheet macro was treated as a defect in the user’s own work; a rogue browser extension was assimilated to traditional intermediary-liability doctrine; a no-code workflow that produced harm was analyzed as a configuration choice for which the user bore responsibility. These

accommodations worked because the user's intervention possessed two features that legal doctrine reliably presupposes: it was *traceable*—one could, if necessary, recover the chain of operations from input to output—and it was *deterministic*—the same input produced the same output, so that the relationship between user choice and product behavior was analytically transparent.

AI-extensible products break both conditions. The user's prompt is not traceable in its causal contribution to the output: the model's processing runs through billions of parameters whose attention patterns and residual streams cannot be reduced to a clean attribution of credit or blame to particular inputs. The output is not deterministic: temperature settings, sampling stochasticity, retrieval-augmented context drawn from other users' recent interactions, and continuous fine-tuning all introduce non-replicable elements. These are not engineering shortcomings to be resolved by better software; they are constitutive features of the generative-AI paradigm itself. Mechanistic interpretability research is making genuine progress, but its current and foreseeable products do not restore the kind of clean traceability on which legal doctrine has historically relied.

The two breaks generate, in turn, the three structural dilemmas that organize Part II. Non-traceability gives rise to *black-box causation*, which destabilizes any doctrine—criminal causation, tort proximate cause, copyright contribution—that requires causal proof to a defined standard. Non-determinism, combined with the multi-user character of contemporary AI services, gives rise to *mixed culpability*: prompts from different users with heterogeneous mental states aggregate, through the model's processing, into a single output whose authorship cannot be cleanly partitioned. The combination, finally, gives rise to *cross-agent liability allocation*: with provider, users, and the AI system itself all implicated and none cleanly attributable, the question of how to allocate responsibility across classes of agent becomes structurally rather than merely incidentally hard. Each of these dilemmas appears, in homologous form, across multiple legal doctrines. That homology—and the structural sources behind it—is the central object of analysis in the chapters that follow.

3. Analytical Case Studies and the Dilemma Matrix

3.1 Methodological Preliminaries

The case studies that follow are constructed rather than drawn from adjudicated facts. They participate in a long methodological tradition—the German *Fallkonstellation*, the American legal hypothetical, the Japanese 設例—in which the analyst stipulates a stylized fact pattern in order to isolate a doctrinal variable for examination. The constructed case is to legal analysis what the controlled experiment is to empirical science: an instrument for separating signal from noise. Real cases bring with them a thicket of incidental detail—procedural posture, evidentiary contingency, unrelated counts, settlement pressure—through which the doctrinal question of interest can rarely be cleanly extracted.

This article uses five constructed scenarios, labelled Scenario α through Scenario ε . Each is designed to isolate a specific configuration of the variables that AI-extensible product harm presents: the number of users involved, the heterogeneity of their mental states, the role of the provider, and the geographic distribution of conduct and consequence. The progression is cumulative: Scenario α fixes the single-user intentional baseline, Scenario β adds the foreseeability problem of single-user

negligence, Scenario γ introduces multi-user mixed culpability, Scenario δ adds cross-jurisdictional dispersion, and Scenario ε —already presented in detail in §3.6—combines all of the above with maximal procedural complexity.

Each scenario is presented in a uniform format: a brief statement of facts and actors, identification of the output and the resulting harm, and an enumeration of the principal doctrinal questions raised across the seven legal fields the article surveys. The treatment is deliberately compact. Detailed doctrinal analysis is deferred to Chapters 4 and 5; this chapter establishes the engine that those chapters then drive.

The reader should not expect any of the constructed scenarios to correspond exactly to a litigated dispute. The closest real-world analogues—the early generative-AI defamation suits, the AI image copyright cases, the deepfake fraud prosecutions—are useful as sources of texture, and are referenced where appropriate, but they present their facts in mixtures that obscure the doctrinal variables this article aims to isolate. The constructed case is the appropriate methodological vehicle when the technology is new and the litigation is sparse; its limits are the limits of analytical legal scholarship generally, and are addressed in Chapter 9.

3.2 Scenario α : Single-User Intentional Misuse

Facts

eA is a domestic user of *Hearth*, the family-management AI-extensible product introduced in §3.1 and Chapter 1. She decides to defraud her elderly aunt, V, of a substantial sum of money by impersonating a trusted family member through a series of text messages. Over the course of an afternoon, eA prompts the product to draft, refine, and personalize a sequence of messages calibrated to V's known vulnerabilities (the recent death of her husband, her concern for a nephew's medical bills, her unfamiliarity with electronic banking). eA reviews each draft, makes minor edits, and dispatches the messages from a phone number she has set up to mimic the nephew's. The fraud succeeds; V transfers the requested sum.

eA's mental state throughout is one of full intention and knowledge of unlawfulness. She has no factual mistake about the identity of the victim, the falsity of the representations, or the unlawful character of her plan. The provider, Company B, has no particular knowledge of eA's use; its automated abuse-detection systems were not triggered by any individual prompt taken in isolation. eA and V are nationals of and resident in the same jurisdiction.

Doctrinal questions raised

Scenario α is the article's baseline case: the configuration in which existing doctrine handles AI-mediated harm with relatively little strain. *Criminal law* reaches eA via indirect perpetration (*mittelbare Täterschaft*, 間接正犯) using the AI as a tool, or—more straightforwardly—via the constituent fraud offense itself (詐欺罪, *Betrug*, wire fraud), with the AI treated as the instrumentality through which the deceptive communication was conveyed. The model's status as a stochastic process introduces no doctrinal innovation, because eA has selected among its outputs and adopted the dispatched messages as her own. *Tort law* reaches her through ordinary intentional-tort doctrine (deceit, conver-

sion of the transferred funds), with the AI again as instrumentality. *Contract law* permits Company B to invoke ToS provisions that prohibit misuse and assign risk to the offending user. *Data protection* adds a layer if the prompts contained personal information about V (which they almost certainly did), but the layer is doctrinally familiar.

The provider's position deserves separate note. Company B attracts no direct primary liability and, on the facts as stated, no secondary liability either: a single user's misuse does not establish the constructive knowledge that contributory liability typically requires, and the product's general functionality (drafting messages) is not inherently unlawful. *Administrative* obligations may apply if the deployment falls within the high-risk tier of the EU AI Act or analogous regimes, but those obligations operate at the level of process compliance rather than per-incident attribution. *Private international law* questions are absent because the configuration is mono-jurisdictional.

Scenario α accordingly serves as the analytical reference point against which the strain on doctrine in subsequent scenarios is measured. Where Scenario α works, the doctrinal machinery is functioning; where later scenarios fail to produce comparably stable answers, the failure is attributable to a feature Scenario α lacks.

3.3 Scenario β : Single-User Negligent or Innocent Use

Facts

eB is a freelance security researcher who uses *Hearth*'s extensible drafting capability for a legitimate professional purpose: documenting the kinds of social-engineering content that elder-fraud schemes have begun to use, with the intention of publishing a defensive write-up for a consumer-protection journal. To produce realistic examples, she prompts the system to generate "a persuasive but generic message that a fraudster impersonating a relative might send to an elderly victim." She receives a generated draft, copies it into a text file on her workstation along with several variants, and posts the file—suitably anonymized and labelled as research material—to a public security-research forum.

A reader of the forum, X, is in fact a fraud practitioner. He adapts eB's research material, with minor edits, into an operational scheme that defrauds V of a substantial sum. eB learns of the use months later from press coverage. Her professional position is intermediate: she has a security-research background and several years of public engagement with social engineering, but she is not an AI-systems specialist and the forum on which she posted is general-purpose rather than adversarial.

Doctrinal questions raised

Scenario β introduces the foreseeability problem at the level of a single user. eB's mental state is at most negligent: she did not intend that her material be used for fraud, and the use by X was, on at least one reasonable view, an intervening criminal act of a third party. But foreseeability admits of degrees, and the precise calibration of the standard of care is contested.

Criminal law in most jurisdictions surveyed declines to extend fraud liability to the negligent contributor. As discussed in §3.6, fraud offenses are everywhere intentional; *fahrlässiger Betrug* is not an offense in Germany, 過失詐欺 is not an offense in Japan, and the equivalent position obtains in the United States, the United Kingdom, China, and the EU member states. The closest available

criminal hooks are generalized aiding offenses (China’s 帮助信息网络犯罪活动罪, Art. 287-2 of the Criminal Law) or, in some jurisdictions, sectoral offenses keyed to the publication of harmful material. Whether eB’s research-purposed publication satisfies their elements is a contested question in each.

Tort law is more demanding of eB. The duty of care owed by a security researcher publishing potentially dangerous material has been litigated, in adjacent contexts, on both sides of the foreseeability question: cases such as *Forsyth v. Eli Lilly* and the German *Lehmann-Brüder* jurisprudence on hacker-tool publication illustrate the range. The result will turn on the jurisdiction’s calibration of the negligence standard, the characterization of eB’s professional position, and the sufficiency of the warnings and anonymization she applied.

Intellectual property questions arise if X’s adapted message is itself a derivative work: who is its author—eB, the AI, or X? *Data protection* questions arise if any of eB’s experimental prompts, or X’s adapted output, contained or generated personal information about identifiable individuals.

The doctrinal stress in Scenario β is concentrated on the foreseeability axis. The same facts will produce different outcomes under different jurisdictions’ calibrations of negligence—a variation that becomes especially consequential in the cross-jurisdictional configurations introduced below.

3.4 Scenario γ : Multi-User with Mixed Culpability

Facts

The product is a productivity assistant deployed by Company B that maintains a shared retrieval-augmented context: prompts and outputs from recent users in a given organizational tenancy or topic cluster influence the model’s responses to subsequent users. Within a four-hour window on a single afternoon, four users in the same jurisdiction interact with the system.

eA prompts the assistant to draft a persuasive financial-help message ostensibly from a relative in distress, intending to use it to defraud her elderly aunt V. She knows the message will be deceptive and intends to use it for fraud.

eB is a marketing professional who, in the same hour, submits prompts asking the assistant to “soften” a fundraising appeal her charity is preparing. Her prompts include phrasing patterns that would be effective in eliciting urgency and emotional commitment. She has no fraudulent intent; her use is for ordinary nonprofit communication. A careful professional in her position might foresee that prompt patterns optimized for emotional urgency could be misused, but she does not turn her mind to the question.

eC is a high-school student preparing an English-class essay on persuasive writing. She prompts the assistant for examples of techniques used in successful direct-mail campaigns. Her use is entirely innocent and there is no plausible basis on which she could foresee any downstream harm.

eD is a member of the same fraud ring as eA, working independently to refine prompt templates for elder-fraud campaigns. He and eA have coordinated only loosely: through pseudonymous accounts on a forum frequented by fraud practitioners, sharing techniques but not specific targets. He shares eA’s general fraudulent purpose but has no knowledge of V or of eA’s specific plan.

The output E —the message eA ultimately receives and sends to V —is shaped by all four users' inputs through the model's context aggregation. eB 's softening patterns supply emotional register, eC 's examples supply structural elements, and eD 's prompt-engineering refinements increase persuasive force. No single user's prompt alone would have produced E ; the harm arises from the cumulative aggregation of contributions through the model's opaque processing. V transfers a substantial sum.

Doctrinal questions raised

Scenario γ is the article's central case study. It exhibits the core multi-user problem in its purest form: four contributors with four different mental states, whose conduct is causally combined through a process whose internal allocation of credit is structurally opaque.

Criminal law confronts a layered attribution problem. eA and eD both acted with intent; both committed conduct that contributed to the deceptive output. But they are not co-principals in the traditional sense, because they shared no common plan with respect to V or to E ; and they are not merely concurrent independent actors, because their contributions interacted through the model. The doctrinal categories of complicity—intentional concert, *Mittäterschaft*, joint enterprise, conspiracy—presuppose either a common plan or essentially identical contributions; neither condition is satisfied here. eB 's negligence is not captured by any applicable fraud offense (which is everywhere intentional), so she escapes criminal liability despite a contribution morally less innocent than eC 's. eC acted innocently and is plainly not criminally liable, but her contribution is causally on a par with eB 's. The doctrine sorts the four actors by mental state, but the sorting produces a result—two convictions, two non-prosecutions—that may seem disproportionate to the relative weights of their respective causal contributions.

Tort law confronts a joint-tortfeasance problem under heterogeneous mental states. The intentional contributors are liable; the innocent contributor is not; the negligent contributor's liability turns on the jurisdiction's calibration of duty and foreseeability. The further question—how to apportion damages among the liable parties—meets the black-box causation problem head-on, because the apportionment cannot be grounded in any defensible quantification of each defendant's contribution to E .

Intellectual property confronts an authorship problem. E was shaped by four creative contributions and an opaque generative process. Whether copyright subsists at all (and if so, in whom) cannot be resolved by the standard joint-authorship test, which requires both intent to merge contributions into a unitary whole and a non-trivial expressive contribution from each putative co-author. *Data protection* confronts a collective-exposure problem if any of the prompts or outputs contained or implicated personal information about V or others.

Administrative regulation raises a question peculiar to the provider: did Company B have a duty to anticipate the aggregation effect—that prompts from different users could combine through shared context to produce harms—and to design the product to prevent it? The EU AI Act's risk-management obligations, the China generative-AI rules, and analogous regimes can be read to impose something like such a duty, but none has yet been authoritatively applied to the multi-user aggregation problem.

Scenario γ is the case in which the article's three structural concepts—black-box causation, mixed culpability, cross-agent liability—all bind simultaneously. It is the central problematic that Chapter 4

analyzes in cross-doctrinal generality and that Chapter 5 traces through each legal field.

3.5 Scenario δ : Cross-Jurisdictional Multi-User Configuration

Facts

Scenario δ adopts the structure of Scenario γ but disperses its actors across borders. The four users—an intentional fraudster, a negligent professional, an innocent contributor, and a second intentional fraudster—are nationals and residents of, respectively, Japan, Germany, China, and Brazil. The provider Company B is incorporated in the United States. Its serving infrastructure is in Singapore. The victim V is resident in the United Kingdom. The harmful output E is generated, dispatched, and acted upon over a period that spans the time zones of all seven jurisdictions.

The substantive misconduct mirrors Scenario γ : cumulative aggregation of contributions through the model's opaque processing produces a deceptive message that defrauds V. The mental states of the four users are as in Scenario γ ; no facts germane to substantive doctrine differ. What differs is the geographic distribution.

Doctrinal questions raised

Scenario δ adds the cross-jurisdictional dimension to the multi-user problem. Each of the doctrinal axes that operated in Scenario γ now acquires a private-international-law overlay.

Criminal jurisdiction can be asserted by each of the home states of the four users (active personality), by the United Kingdom (territorial, passive personality, protective principle if the fraud is characterized as targeting a UK resident), and arguably by the United States (provider connection, wire-fraud reach). *Extradition* is constrained by the German, Brazilian, and Chinese constitutional or statutory bars on extraditing nationals; only the US fraudster is reliably reachable through extradition. The remaining prosecutions, if they occur at all, occur in the home states under *aut dedere aut judicare*, with the well-known coordination deficits.

Choice of law in tort engages Rome II for any EU forum, the Hague Convention on the Law Applicable to Products Liability for the limited set of states party to it, and traditional common-law analysis elsewhere. The result is plural: a UK action against Company B may apply English law as the law of the place of damage, while a German action against the German user may apply German law, and a Brazilian counter-claim may apply Brazilian law. *Choice of law in copyright* engages the Berne Convention's *lex loci protectionis* principle, generating the prospect of parallel rights determinations under each national copyright regime. *Data protection* engages GDPR (for the German user and any EU processing), PIPL (for the Chinese user and any China processing), APPI (for the Japanese user), the LGPD (for the Brazilian user), and the UK GDPR (for the victim). *Administrative regulation* of the provider is contested between the EU AI Act (extraterritorial under Art. 2 if outputs are used in the EU), the UK Online Safety Act, China's generative-AI rules (if Company B provides services in China), and the patchwork of US sectoral regimes.

The practical result, generally observed across this category of cases, is what may be called *lowest-hanging-fruit enforcement*: of the multiple plausible cases, only those in which jurisdictional reach, evidentiary feasibility, extradition treaty coverage, and prosecutorial capacity converge will actually

be brought. This is typically the intentional user in a well-integrated legal system (here, the US fraudster), or—failing that—a civil action by the victim in her home forum against whichever defendant is most easily reachable through service of process. The negligent and innocent contributors, the foreign provider, and the AI system itself drop out of the practical attribution map regardless of their substantive doctrinal position. Scenario ε , presented in detail at §3.6, traces this enforcement reality through each procedural layer.

3.6 Scenario ε : A Six-Actor Multinational Configuration with Heterogeneous Culpability

Where Scenario δ introduces the cross-jurisdictional dimension at the level of a four-actor configuration, this scenario extends both the number of actors and the heterogeneity of their mental states, and traces the resulting case through the entire procedural pipeline from jurisdictional assertion to judgment enforcement. The construction is deliberately maximal: it is unlikely that any single real-world fact pattern will present every difficulty rehearsed here, but each difficulty arises naturally in at least some realistic permutation of AI-extensible product misuse.

3.6.1 Facts

The product is an AI-extensible communication assistant deployed by Company B, a company incorporated in Delaware (United States), with its European establishment registered in Dublin (Ireland), and its serving infrastructure located in data centers in Singapore. The system processes user prompts through a generative model that maintains short-term context across recent user inputs and incorporates retrieval-augmented responses drawn from a continuously updated corpus.

Six users interact with the system over a 48-hour period.

- **eA (national: Japan; physically located: Tokyo)** submits prompts with full knowledge that she intends to use the generated outputs to defraud an elderly relative of her employer's business partner. She has researched the victim, prepared a script, and uses the AI to refine and personalize the deceptive communication.
- **eB (national: Germany; located: Berlin)**, a software engineer with several years of professional experience in AI systems, posts a sequence of prompt-engineering experiments to a public technical forum, demonstrating "how to elicit more persuasive persuasive-writing outputs." He intends only to share technical findings. The prompt templates he publishes are subsequently incorporated into eA's workflow.
- **eC (national: China; located: Shanghai)** submits a series of prompts that materially shape the model's output. The prompts are ambiguous in purpose: some are consistent with innocent use (writing assistance for a personal letter), others are consistent with deliberate refinement of a deceptive script. The evidentiary record will prove insufficient to establish her subjective state of mind beyond reasonable doubt.
- **eD (national: United States; located: San Francisco)** coordinates with eA indirectly: the two communicate only through pseudonymous accounts on a forum frequented by fraud practitioners, and have never identified themselves to each other. eD's prompts are intentionally crafted to support

the same fraudulent campaign eA is conducting. He is fully aware of the unlawful purpose.

- **eE (national: United Kingdom; located: London)** submits a single, unrelated query about persuasive writing for a high-school essay. By coincidence of context aggregation, her prompt influences the retrieval-augmented context within which eA's final output is generated. She has no knowledge of any fraudulent activity and no reasonable basis to suspect any.

- **eF (national: Brazil; located: São Paulo)** is a freelance journalist who has been told by eA (through a fabricated cover story) that the prompts she is asked to test are part of a research project on elder financial vulnerability. Her contributions materially shape the output, but she labors under a fundamental mistake about the nature of the work she is assisting.

The victim, **V**, is a Korean national resident in Australia. The fraudulent communication is delivered to her in Australia, and the financial harm crystallizes when she transfers funds from her Australian bank account to an offshore account.

The output *E* that produces the harm is generated by the interaction of inputs from all six users. No single user's contribution is, in isolation, sufficient to produce *E*; the harm arises from the cumulative aggregation of contributions through the AI system's opaque processing.

3.6.2 The Procedural Pipeline

A criminal proceeding against any subset of eA through eF must pass through seven procedural layers, each of which may halt or constrain the case independently of the others. We sketch each in turn before proceeding to actor-by-actor analysis.

Jurisdictional assertion. At least eight jurisdictions can plausibly assert criminal jurisdiction over some portion of the facts: Japan, Germany, China, the United States, the United Kingdom, and Brazil (each on the basis of active personality or territorial conduct by their nationals); Korea (passive personality, since the victim is a Korean national, under Korean Criminal Code Art. 6); and Australia (territorial, since the harm crystallized there). Additional bases exist for the EU under the AI Act's extraterritorial provisions (Art. 2) targeting the provider rather than the users.

Investigation and evidence gathering. Because the relevant evidence is distributed across users' devices, the provider's servers (Singapore), and the financial intermediaries that processed the transfers, no single jurisdiction can develop a complete evidentiary record without international cooperation. The Council of Europe Cybercrime Convention (Budapest Convention) provides a framework to which Japan, Germany, the United Kingdom, the United States, Australia, and Brazil are all party; China and Korea are not. Mutual legal assistance treaties exist bilaterally among many of the jurisdictions involved, but the China–Australia and China–United States channels are slow and politically constrained.

Prosecution coordination. No global authority allocates prosecutorial priority among interested states. Within the European Union, Eurojust performs this function for member states; for the broader configuration here, prosecutors must coordinate informally. In practice, the first jurisdiction to indict

typically establishes lead status, and others either join (cooperatively, through MLAT) or stand down (deferring on grounds of comity, or because the lead's prosecution will plausibly satisfy the prosecutorial interest). Australia, as the locus of harm and with most of the documentary evidence (financial records, the victim's testimony) within its territory, is the most likely lead.

Personal jurisdiction over the defendant. A criminal trial typically requires the defendant's presence (with exceptions for *in absentia* proceedings in some civil-law jurisdictions). The defendant's presence requires either voluntary appearance or extradition. Extradition is governed by bilateral treaties and is subject to several universally recognized exceptions:

- **Non-extradition of nationals.** Germany (Grundgesetz Art. 16), Brazil (Constitution Art. 5, LI), and China (Extradition Law Art. 8) categorically refuse to extradite their own nationals. Japan does not have a constitutional prohibition but extradites its nationals only on a discretionary basis under the Law of Extradition Art. 2, and in practice rarely does so.

- **Dual criminality.** The conduct must be criminal in both the requesting and the requested state.

- **Specialty.** The defendant may be tried only for the offenses for which extradition was granted.

- **Political offense exception, death penalty exception, human rights review.**

Under *aut dedere aut judicare* (extradite or prosecute), a state that refuses extradition is generally expected to prosecute the offense itself if it has jurisdiction. The expectation is normative and inconsistently honored.

Applicable substantive law. Each forum applies its own substantive criminal law. An Australian prosecution applies the Australian CRIMINAL CODE ACT 1995 (particularly Ch. 7 on offenses involving dishonesty); a Japanese prosecution applies the PENAL CODE (Art. 246 on 詐欺罪); a German prosecution applies StGB (§263 on *Betrug*); and so on. Differences in the constitutive elements of comparable offenses (notably the foreseeability standard for negligent variants, where they exist) can produce different outcomes on the same facts.

Judgment and execution. A conviction in one jurisdiction must be enforced there or, by transfer of sentenced persons under applicable treaties, in another. For defendants who never enter the convicting jurisdiction, an *in absentia* conviction is enforceable only in jurisdictions that recognize the proceeding. Recognition of foreign criminal judgments is substantially more limited than recognition of civil judgments under treaties such as the Hague Convention of 2 July 2019.

Ne bis in idem. Within the European Union, Art. 54 of the Schengen Implementation Convention provides a robust transnational *ne bis in idem* protection: a person finally judged in one Schengen state may not be prosecuted in another for the same acts. No equivalent global provision exists; a defendant convicted or acquitted in Japan could in principle be prosecuted again in Australia for the same conduct, subject only to domestic-law constraints in the second jurisdiction.

802 3.6.3 Actor-by-Actor Analysis

803 We apply this pipeline to each of the six actors in turn. The analysis proceeds on the working as-
 804 sumption that prosecution authorities would, under the constraints sketched above, prosecute where
 805 conviction is realistically achievable.

806 **eA (Japanese national, intentional misuse).** Substantively, eA's conduct satisfies the elements of
 807 indirect perpetration (間接正犯) using the AI system as a tool, or alternatively the elements of 詐欺
 808 罪 (Penal Code Art. 246) on the theory that the model's output is the means by which the deception
 809 is conveyed. Jurisdictionally, Japan has territorial jurisdiction (the acts occurred in Tokyo) and Ko-
 810 rea, Australia, and the United States (through the provider connection) can also assert jurisdiction.
 811 Procedurally, Japan will not extradite eA. Australia, recognizing this, will likely defer to a Japanese
 812 prosecution under *aut dedere aut judicare*, providing evidentiary assistance under the Japan–Australia
 813 MLAT. The likely outcome is conviction in Japan, with a custodial sentence in the range typical for
 814 organized fraud (3–10 years depending on aggravating factors).

815 **eB (German national, gross negligence).** Substantively, eB's conduct may amount to gross neg-
 816 ligence: a professional in AI systems posting prompt techniques publicly should foresee, perhaps,
 817 that such techniques could be misused. But neither German law nor the laws of the other interested
 818 jurisdictions provides for a negligent variant of fraud. §263 StGB punishes intentional *Betrug* only;
 819 *fahrlässiger Betrug* is not an offense. The same is true under Japanese, Australian, Korean, and
 820 Chinese law for their respective fraud provisions. Absent a statutory negligent offense, eB is not
 821 criminally liable anywhere, despite his contribution being morally blameworthy and civilly action-
 822 able. Civil liability under §823 BGB (German law of torts) is plausible if causation can be shown to
 823 a civil standard. The prosecutorial reality: eB is not charged. A civil action by V against eB may be
 824 brought in Germany under the law applicable according to Rome II.

825 **eC (Chinese national, undetermined intent).** Substantively, the principal difficulty is evidentiary.
 826 Chinese Criminal Procedure Law Art. 53 requires that conviction rest on evidence that excludes rea-
 827 sonable doubt. If eC's subjective state of mind cannot be established to that standard, the principal
 828 fraud charge fails. But Chinese criminal law offers a doctrinally narrower but evidentially easier
 829 alternative: the offense of aiding information network crime (帮助信息网络犯罪活动罪) under
 830 Art. 287-2 of the Criminal Law, introduced in the 2015 amendment. The Two Highs' Judicial Inter-
 831 pretation of 2019 lowers the threshold for “明知” (knowledge) in this offense by permitting inference
 832 from objective factors. A prosecution under Art. 287-2 in China is realistically possible; a prosecution
 833 under the main fraud offense (诈骗罪, Art. 266) is not. Jurisdictionally, China has territorial juris-
 834 diction. Procedurally, China will not extradite eC; if Australia or Korea wishes to see her prosecuted,
 835 only Chinese authorities can do so. The outcome depends on political and prosecutorial choices that
 836 are opaque to predictability; in many similar past cases, such prosecutions have not been pursued
 837 absent political salience.

eD (US national, intentional, coordinating with eA). Substantively, eD's conduct readily satisfies the elements of US wire fraud (18 USC §1343), which has broad extraterritorial reach where wires or electronic communications cross the US border. Jurisdictionally, the United States, Australia, Korea, and Japan all have plausible bases. Crucially, the United States does extradite its nationals: the Extradition Treaty between the United States and Australia of 1974 (amended 1990) imposes no nationality bar, and Australia could request eD's extradition. Equally, the United States would likely prosecute eD domestically under §1343. The doctrinal question of whether eD and eA are co-conspirators or merely concurrent intentional actors turns on the sufficiency of the forum-based coordination as evidence of agreement; under *Rosemond v. United States* (572 U.S. 65 (2014)), advance knowledge of all elements is required for accomplice liability. The pseudonymous forum communication may or may not satisfy this. Conviction as a sole intentional perpetrator using AI as a tool is, however, secure on the facts as stated. The likely outcome: prosecution in the United States, conviction, custodial sentence.

eE (UK national, no fault). Substantively, eE lacks both *mens rea* and any wrongful *actus reus*: her prompt was innocent on its face and at the time of submission. Even under common-law strict liability or status offenses, no element of any relevant offense applies to her conduct. The fact that her input causally contributed to *E*, through unforeseeable context aggregation, does not create criminal liability in the absence of fault. Civilly, the causation chain is too remote and the lack of fault too clear for tort liability. No jurisdiction will prosecute eE; no civil claim against her would succeed. She is a casualty of the black-box causation problem in its purest form: an entirely innocent party whose conduct nevertheless figured in the causal chain leading to harm.

eF (Brazilian national, mistake of fact). Substantively, eF's case turns on the doctrine of mistake of fact (*erro de tipo*, Brazilian Penal Code Art. 20; *Tatbestandsirrtum* in German doctrine; mistake of fact generally in common law). Where her mistake was as to a circumstance constitutive of the offense (the existence of fraudulent purpose), the mistake negates the *mens rea* of any intentional offense. The mistake itself was non-culpable: she was actively deceived by eA through a credible cover story, and her reliance on the cover story was reasonable for a freelance journalist. The mistake therefore precludes conviction for an intentional offense; absent a statutory negligent variant (which, as noted, fraud lacks in all jurisdictions surveyed), no conviction is available. Jurisdictionally, Brazil has territorial jurisdiction; Brazil will not extradite eF. Outcome: non-prosecution, or prosecution and acquittal on the basis of mistake of fact.

3.6.4 The Enforcement Reality

The combined effect of these procedural and substantive constraints produces an outcome whose distributive structure deserves attention. Of the six actors:

- Two (eA, eD) are convicted, each in their home jurisdiction, for intentional offenses, receiving substantial custodial sentences.
- Two (eB, eE) face no criminal liability: eB because his negligence is not captured by any applicable offense, eE because she committed no wrongful act.

- 876 • One (eC) faces an outcome contingent on prosecutorial choice, with a plausible conviction under
877 a narrow statutory offense (帮信罪) but acquittal under the main fraud charge.
- 878 • One (eF) is acquitted or not prosecuted on the basis of mistake of fact.

879 This distribution is, in one sense, a vindication of the substantive criminal law's commitment to
880 fault: those who acted intentionally bear the heaviest consequences, those who acted innocently bear
881 none, and the intermediate cases are sorted by their specific doctrinal status. In another sense, the
882 distribution is troubling: the victim V has been defrauded through the combined operation of contri-
883 butions from six actors and the AI system; she will recover, if at all, only against those few defendants
884 who happen to be reachable through civil process in jurisdictions that recognize her standing. The
885 provider Company B, whose product was the medium through which the harm was produced, will
886 face administrative scrutiny under EU AI Act provisions (if the deployment qualifies as high-risk) and
887 possibly civil liability under the revised Product Liability Directive, but is unlikely to face criminal
888 sanction in any jurisdiction, given the application of the Winny standard and its functional equivalents
889 elsewhere.

890 The enforcement of any judgment compounds these difficulties. eA's Japanese conviction is en-
891 forceable in Japan, where she is. eD's United States conviction is enforceable in the United States,
892 where he is, or potentially in Australia if extradition occurred. None of the convictions can be straight-
893 forwardly enforced against assets located in jurisdictions other than the convicting one without further
894 treaty-based recognition processes, most of which are calibrated for civil rather than criminal judg-
895 ments.

896 3.6.5 What Scenario ϵ Reveals

897 Three points emerge from this analysis that the simpler scenarios do not expose with equal clarity.

898 First, *substantive law and procedural law are non-separable in the multinational AI context*. A
899 defendant's culpability under abstract substantive doctrine may bear no relationship to her actual
900 criminal exposure, which is determined as much by extradition treaty coverage, prosecutorial coor-
901 dination, and evidentiary feasibility as by the merits of her case. This is not a defect specific to AI-
902 extensible products; it is a long-standing feature of transnational criminal law. But the AI-extensible
903 product paradigm amplifies the feature dramatically, because such products routinely produce facts
904 that straddle six or more jurisdictions in a single incident.

905 Second, *the doctrine of mistake of fact, the absence of negligent fraud offenses, and the in dubio*
906 *pro reo principle together generate substantial under-enforcement in the AI context*. This is not a
907 normative criticism: these doctrines reflect deep commitments of criminal law to fault-based liability
908 and the protection of the innocent. But the combined effect, in scenarios like the present one, is
909 that intermediate cases of contribution to harm fall outside the criminal law's reach, and must be
910 addressed—if at all—through civil and administrative channels. This reinforces the policy argument,
911 developed in Chapter 8, that ex ante regulation and civil remedies must bear the principal weight of
912 governing AI-extensible product harm.

913 Third, *the cross-border enforcement gap is structurally victim-adverse*. The defendants in this
914 scenario who are most reachable (eD in the United States, eA in Japan with cooperative domestic

prosecution) are those who happen to live in well-integrated legal systems. Defendants in jurisdictions that do not extradite their nationals and do not consistently prosecute (a category that, on the facts of this scenario, includes parts of eC’s and eF’s situations) face significantly less exposure, regardless of substantive culpability. The victim V, accordingly, faces a recovery process in which her access to justice depends substantially on factors orthogonal to the merits of her claim. This is the empirical reality that any diagnostic analysis of AI-extensible product harm must register, and that any prescriptive proposal must engage.

3.7 The Dilemma Matrix

The five scenarios sketched in the preceding sections, mapped against the seven doctrinal fields the article surveys, generate a 5×7 table of dilemmas. The diagnostic value of the table lies less in any particular cell than in the patterns that emerge across rows and columns. Each cell records, in compressed form, the principal dilemma raised by that scenario in that doctrinal field; cross-references direct the reader to the corresponding analysis in Chapters 4 and 5.

The matrix is presented landscape and in compact form for readability. It is not exhaustive: other doctrines (labor law, consumer protection law, competition law, securities law) may be implicated in some cells but are addressed only briefly in §5.8. Nor is the diagnosis in each cell intended as a final answer; it is a flag identifying the principal source of doctrinal strain.

Table 2: Dilemma matrix: scenarios $\times$ doctrinal fields.

	Criminal	Tort	IP	Contract / ToS	Data protection	Administrative	PIL
α (single intentional)	[D] None substantial; indirect perpetration handles it.	[D] None substantial; intentional tort doctrine handles it.	[D] Authorship of <i>E</i> unclear; ToS may resolve.	[D] Enforceability of misuse-prohibition clauses against user.	[D] PII of victim in prompts; ordinary consent issues.	[D] Provider’s risk-management obligations under high-risk regimes.	[D] Mono-jurisdictional; PIL inactive.
β (single negligent)	[D] No negligent fraud offense in any major jurisdiction; criminal liability fails.	[D] Foreseeability standard contested; calibration to user’s professional role.	[D] Authorship of derivative use by third party.	[D] ToS warranty disclaimers; limits under unfair-terms law.	[D] Duty to anticipate inadvertent PII exposure.	[D] Status of user as “employer” under AI Act.	[D] Mono-jurisdictional in stipulated form.

	Criminal	Tort	IP	Contract / ToS	Data protection	Administrative	PIL
γ (multi-user mixed)	[D] Layered attribution; complicity doctrine fails for non-concerted actors.	[D] Joint tortfeasance under heterogeneous mental states; apportionment indeterminate.	[D] Joint authorship test fails; rights uncertain.	[D] ToS cannot allocate among multiple users without notice.	[D] Aggregation effect; collective exposure across prompts.	[D] Provider's duty to anticipate aggregation.	[D] Mono-jurisdictional; deferred to δ .
δ (cross-jurisdictional)	[D] Concurrent jurisdictions; extradition bars; <i>ne bis in idem</i> gaps.	[D] Choice of law (Rome II); plural applicable laws.	[D] <i>Lex loci protectionis</i> ; parallel national copyrights.	[D] Choice-of-forum and choice-of-law clauses' enforceability.	[D] Concurrent regimes (GDPR, PIPL, APPI, LGPD).	[D] Concurrent regulators; AI Act extraterritorial reach.	[D] All PIL questions live; recognition and enforcement gaps.
ε (six-actor multinational)	[D] All of δ plus: mistake of fact (eF); 帮信罪 as alternative (eC); δ 's extradition gaps amplified.	[D] All of δ plus: PLD 2024 against provider; AILD presumptions if adopted.	[D] All of δ plus: divergent treatment of AI authorship across six fora.	[D] All of δ plus: ToS conflicts with mandatory consumer law in multiple fora.	[D] All of δ plus: cross-border transfer mechanisms (SCCs, adequacy).	[D] All of δ plus: AI Act vs. China rules vs. UK OSA on the same provider.	[D] All of δ plus: enforcement asymmetry; victim recovery limited to reachable defendants.

932 3.8 Reading the Matrix: Three Patterns

933 Three patterns are visible in Table 2 that the remainder of the article works through in detail.

934 **Diagonal severity.** The dilemmas intensify as one moves from Scenario α through Scenario ε .
935 Scenario α 's row contains few entries marked as substantial; the doctrinal machinery handles the single
936 intentional case competently across most fields. Scenario β 's row introduces the foreseeability
937 problem and the limits of negligent-variant offenses. Scenario γ 's row makes layered attribution and
938 joint authorship live across most columns. Scenario δ 's row activates every PIL cell and multiplies
939 most others. Scenario ε 's row contains the maximal configuration. The progression is not coincidental:
940 the case studies are constructed to vary the salient dimensions monotonically, so that the increase
941 in dilemmatic load can be attributed to the introduction of specific features (multiplicity, mixed culpability,
942 jurisdictional dispersion). This pattern motivates the case-study framework as the analytical
943 engine of the article.

944 **Doctrinal homology.** Reading down the columns rather than across the rows reveals that similar
945 structural problems—causation under opacity, attribution across heterogeneous contributors, intermediary
946 liability under non-deterministic generation—reappear in cells across multiple doctrines.
947 Black-box causation generates parallel difficulties in criminal proof, tort proximate cause, and copy-
948 right contribution analysis. Mixed culpability disrupts complicity in criminal law, joint tortfeasance

in tort, and joint authorship in IP. Cross-agent allocation runs through criminal indirect-perpetration doctrine, tort joint-and-several rules, and the safe-harbor regimes of contract and administrative law. The homology suggests that the dilemmas are not idiosyncratic to particular doctrines but trace to structural features of the underlying paradigm. Chapter 4 develops this claim.

Doctrinal specificity. A subset of dilemmas is, by contrast, genuinely doctrine-specific and does not reduce to the three structural sources. The authorship question in copyright (whether AI-only contributions can be authored at all, or whether they fall into a “no-rights” zone) is not a causation problem; it is a problem about the personhood predicate of authorship that has no obvious analog in the criminal or tort columns. The enforceability of ToS provisions under unfair-contract-terms law turns on consumer-protection doctrines that are themselves jurisdiction-specific. The interaction of PIL choice-of-law rules with substantive doctrines varies in ways that cannot be predicted from the substantive doctrines alone. These doctrine-specific difficulties motivate Chapter 5’s field-by-field treatment.

The two organizing chapters of Part II accordingly track the second and third patterns respectively. Chapter 4 examines the homology—black-box causation, mixed culpability, cross-agent liability allocation—across doctrines. Chapter 5 examines the doctrine-specific complications that the structural analysis leaves over.

4. Common Structural Dilemmas: Three Cross-Doctrinal Sources

4.1 Introductory Framing

The previous chapter’s matrix exhibited a pattern this chapter argues to be more than coincidence: the dilemmas raised by AI-extensible products, read across doctrines, share a small number of underlying structural sources. Each source is a feature of the technology–practice complex that AI-extensible products instantiate; each manifests in homologous form in fields as doctrinally distinct as criminal law, tort, copyright, and data protection. The claim of homology is the article’s first theoretical contribution, and the present chapter develops it.

Three sources are central. *Black-box causation* is the structural unknowability of the causal contribution of a particular input to the output of a contemporary generative AI system. It is distinct from, and more radical than, the familiar problem of scientifically unascertained causation that has preoccupied courts in mass-tort and environmental contexts. *Mixed culpability* is the configuration in which multiple agents contribute to a single outcome through inputs reflecting heterogeneous mental states; existing complicity and joint-tortfeasance doctrines presuppose either intentional concert or fundamentally similar contributions, and the mixed-culpability scenario satisfies neither. *Cross-agent liability allocation* is the unsettled distribution of responsibility among product provider, users, and the AI system itself; the doctrine offers fragmentary tools but no coherent framework integrating them.

These three sources do not exhaust the set of difficulties AI-extensible products generate, and a more extensive analysis might identify a fourth or fifth. They are, however, the principal sources of doctrinal failure across the cases analyzed in Chapter 3, and they suffice to organize the analysis

that follows. Each subsection that follows takes one source in turn: traces its appearance across multiple doctrines, identifies the existing doctrinal tools that engage it most directly, and shows why those tools are inadequate to the task the AI-extensible product paradigm presents. §4.5 draws the homology together and bridges to the doctrine-by-doctrine analysis of Chapter 5.

4.2 Black-Box Causation

4.2.1 Traditional Causation and Its Epistemological Premise

The principal causation doctrines of modern legal systems share an epistemological premise that is rarely made explicit. The *conditio sine qua non* (but-for) test, the doctrines of adequate and proximate causation, the German theory of objective imputation (*objektive Zurechnung*), and the American Restatement (Third) of Torts' factual-causation framework all presuppose that the causal chain from conduct to harm is, in principle, ascertainable. The presupposition operates at two levels: that the underlying physical or social mechanism is in principle knowable to science, and that the legal proceeding can, with appropriate evidence, arrive at a defensible reconstruction of it.

Hart and Honoré's *Causation in the Law*, still the canonical treatment, makes this presupposition central. Their account distinguishes between cases in which causation is established and the attribution question is normative (the proximate-cause inquiry) and cases in which causation itself is contested. In both, the framework assumes that the causal facts are facts—that the question of what caused what is, in principle, answerable—even where the law's threshold of proof leaves some questions unresolved in the particular case. Roxin's elaboration of objective imputation in the German criminal-law tradition adds normative filters but preserves the same underlying ontology: conduct can be objectively attributed to a result when, among other things, the conduct created or increased a relevant risk that materialized in the result. The risk-creation analysis requires, in its turn, a tractable model of the causal mechanism.

The American Restatement (Third) of Torts §26 codifies factual causation as a counterfactual: harm is factually caused by conduct if the harm would not have occurred without the conduct. The counterfactual is meaningful because we suppose ourselves capable, at least with sufficient information, of evaluating the relevant counterfactual world. The Restatement treats the rare cases of genuine ascertainment failure as limit cases, addressed through substantial-factor doctrines, market-share liability, and loss-of-chance frameworks; these are exceptions that preserve the rule.

Across these doctrinal families, the epistemological premise is the same. Where the premise holds, doctrine functions: the difficult questions are normative (which causes are legally relevant?), and the factual questions are tractable. Where the premise fails, doctrine strains. The strain is the subject of the next two subsections.

4.2.2 Scientific Uncertainty as the Existing Exception

Legal systems have, over the past half-century, developed an elaborate secondary apparatus for cases in which the underlying scientific evidence cannot conclusively establish causation to the standard the substantive doctrine requires. The mass-tort and environmental-harm contexts are the principal sites of this development. Japanese courts in the toxic-injury jurisprudence—the Morinaga arsenic

milk case, the Minamata methylmercury cases, the SMON drug-injury cases—worked out a doctrine of 疫学的因果関係 (epidemiological causation) that permits liability to be grounded in statistical association where direct biochemical pathways are unprovable on the individual case. German courts have applied analogous probabilistic reasoning under the rubric of *Anscheinsbeweis* and, in product-liability contexts, statutory presumptions that shift evidentiary burdens to the defendant. American jurisprudence, after *Daubert v. Merrell Dow Pharmaceuticals*, has built a substantial body of expert-testimony admissibility doctrine that adjudicates whether particular causal claims are sufficiently grounded in scientific methodology to be presented to a jury; the substantive doctrines of market-share liability (*Sindell v. Abbott Laboratories*), loss-of-chance (*Herskovits v. Group Health Cooperative*), and toxic-tort double-the-rate inferences address particular configurations where individual causation is unprovable.

What unites this apparatus is the shared assumption that the underlying question—which exposure caused which harm—is, in principle, a question to which science could give an answer if it had sufficient data and time. The legal doctrines respond to a temporary or practical limitation on knowledge, not a categorical one. In the DES cases, market-share liability was justified precisely because each plaintiff’s mother had ingested a particular manufacturer’s product, and that fact was simply unrecoverable; not because the question of which manufacturer was inherently meaningless. In epidemiological causation, the legal acceptance of probabilistic proof rests on the assumption that the population-level association is a proxy for an individual-level causal mechanism that we cannot directly observe.

The exceptions, in other words, are calibrated to the failure mode of ordinary scientific inquiry: data scarcity, evidentiary destruction, mechanism complexity that nonetheless yields to investigation in principle. Whether they are calibrated to a different failure mode—one in which the causal question is not under-investigated but constitutively unanswerable—is the question taken up next.

4.2.3 Why AI Black-Box Causation Is Different

The causal opacity of contemporary generative AI is structurally distinct from the temporary scientific opacity addressed by the mass-tort apparatus. Three features account for the difference, and together they exclude the existing apparatus as a complete response.

First, the model’s architecture is not a black box in the sense that its inner workings are hidden from inspection; its weights and activations are, in a literal sense, fully accessible to anyone with access to the model. The opacity is rather a matter of cognitive intractability: a contemporary frontier model contains hundreds of billions of parameters interacting through dozens of attention layers, and the path from any particular input token to any particular output token traverses a high-dimensional manifold that no human (and no analytic procedure now known) can reduce to a clean attribution of credit. Burrell (2016) distinguished this kind of opacity—arising from the scale and structure of the system—from the opacity of trade secrecy and the opacity of technical illiteracy; the present discussion concerns Burrell’s third type. Lipton (2018) has further argued that the very notion of model interpretability is contested, with multiple incompatible criteria (transparency, post-hoc explanation, simulability) that cannot all be jointly satisfied for systems of contemporary scale.

Second, mechanistic interpretability research—which has made genuine and accelerating progress

in the past several years—does not, in its current or foreseeable form, recover the kind of input-to-output attribution that legal causal analysis requires. The methods that have proven most productive (sparse autoencoders for feature extraction, attention-pattern analysis, activation patching) generate approximate, qualitative accounts of what particular model components appear to do; they do not produce per-output decompositions of the kind needed to say, with the confidence required for legal proof, that input x_1 contributed w_1 to output y and input x_2 contributed w_2 . Rudin (2019) has argued that interpretable models should be preferred for high-stakes applications precisely because post-hoc interpretation of opaque models is structurally limited; the argument carries directly into the legal-causation context.

Third, the training-data substrate is not retrievable from the trained model. The information that the model has absorbed during training is encoded in its parameters in distributed and non-localized form; no procedure now available extracts which training documents materially shaped a given output. The result is that the model's output is causally influenced by sources (the training corpus) and processes (gradient updates over many epochs) that are, after the fact, fundamentally inaccessible to investigation in the way that ingredient lists, exposure manifests, and process documentation are accessible in the mass-tort cases.

A concrete illustration may help. A transformer language model generating an output token computes attention weights from the current context to itself across multiple heads and layers. The token finally emitted is the result of a sampling operation over a probability distribution that itself reflects the cumulative product of these weights, applied to value vectors derived from prior representations. Suppose three users have submitted prompts that have entered the model's context (perhaps via retrieval augmentation). To attribute the model's choice of a particular output token to a specific user's prompt requires identifying which attention paths through which heads and layers carried information from that prompt to the final distribution, and what fraction of the final distribution's mass over the chosen token can be assigned to those paths rather than to others. The required computation is well-defined in principle and intractable in practice; even if performed, its result would not be a clean "user i 's prompt was w_i percent responsible" but a multi-dimensional pattern of attention weights that resists reduction to a single attribution number. The standard of proof in any substantive doctrine (beyond reasonable doubt, preponderance of the evidence, even the more permissive "substantial factor") has no purchase on this representation.

The conclusion is that black-box causation in the AI-extensible product context is not a knowledge gap to be closed by more research. It is a constitutive feature of the technology that legal doctrine must engage on its own terms.

4.2.4 Cross-Doctrinal Implications

The structural unknowability of AI causation manifests across multiple doctrines in parallel. The homology is the central observation: the same underlying feature creates analogous strain in criminal, tort, IP, and data-protection settings, and any adequate response must engage the underlying feature rather than each doctrine separately.

In *criminal law*, the standard of proof for causation is the demanding one of beyond reasonable doubt (合理的疑いを超える), and the constitutive elements of fraud, theft-by-deception, and re-

lated offenses require that the accused's conduct caused the harm in a sense that the prosecution can prove. Where the conduct's contribution is only one of multiple inputs aggregated through an opaque generative process, satisfying the standard becomes structurally hard. The same difficulty appears in negligent-result offenses where they exist; the gross-negligence variants in many jurisdictions require proof that the defendant's negligence caused the harm, and the proof problem is not merely difficult but, in the strict sense, unanswerable.

In *tort law*, the preponderance-of-the-evidence standard is nominally more accommodating, but the underlying problem persists. Demonstrating that more likely than not a particular defendant's conduct caused the plaintiff's harm requires some defensible model of the contributory weights of competing causes. Black-box causation denies the model. The market-share liability and loss-of-chance adaptations developed for mass-tort and medical-malpractice contexts were calibrated to specific scientific failure modes (untraceable DES manufacturer; competing causes of cancer with quantifiable base rates); they do not transfer cleanly to the AI-output configuration, where the relevant "market" is not the population of actors but the population of inputs to a single processing event.

In *intellectual property*, particularly copyright, joint-authorship analysis requires assessment of each putative co-author's creative contribution to the work. Black-box causation makes this assessment impossible: each user's prompt is a creative contribution that the model has incorporated, but the share of the output's expressive content attributable to each prompt is structurally indeterminate. The Beijing Internet Court's 2023 AI-image judgment finessed this problem by attributing copyright to the user whose prompt was most proximate to the eventual output; the more general multi-user case that AI-extensible products will increasingly present cannot be so finessed.

In *data protection*, the right of access (GDPR Art. 15), the right to explanation in automated decision-making (Art. 22), and the underlying transparency principle all assume that the data controller can reconstruct, at least approximately, which inputs produced which outputs. Black-box causation falsifies this assumption. Where multiple users' prompts contributed to an output that exposed personal information of a third party, no procedure now available identifies which user's contribution was the proximate exposure; the data subject's remedies are correspondingly thin.

These parallel manifestations are not coincidence. They reflect a shared epistemological structure that contemporary doctrine does not adequately accommodate. The next two structural sources operate in analogous cross-doctrinal fashion.

4.3 Mixed Culpability of Multiple Agents

4.3.1 The Binary Structure of Traditional Complicity

The doctrines of complicity in modern criminal-law systems are organized around a binary distinction. On one side stand the doctrines of intentional concert: co-perpetration (*Mittäterschaft*, 共同正犯, joint principalship), conspiracy and the related common-law joint-enterprise doctrines, the civil-law accomplice and abettor categories. These doctrines presuppose that the participants share, to a sufficient degree, a common plan or mutual awareness of each other's contribution: *gemeinsamer Tatentschluss* in German doctrine, 意思の連絡 in Japanese, the agreement element of conspiracy at common law. The participants need not have identical roles or identical knowledge of details, but the

1144 binding element is some form of subjective convergence.

1145 On the other side stand the doctrines of concurrent independent action: the parallel-tortfeasor cat-
 1146 egories of civil law, the multiple-author fraud doctrines of common law, the special evidentiary pro-
 1147 visions for cases of indistinguishable conduct (the Japanese 同時傷害特例 of Penal Code Art. 207,
 1148 the German *alternative Kausalität* doctrine). These doctrines do not require subjective convergence,
 1149 but they presuppose either that the multiple actors' contributions are essentially similar (each could
 1150 have produced the harm alone, the apportionment is between indistinguishable causes) or that the
 1151 harm is divisible among them on some defensible basis.

1152 The intentional-concert side handles the case where two fraudsters plan together; the concurrent-
 1153 independent side handles the case where two hunters fire simultaneously and one bullet hits a third
 1154 party. Between them, these doctrines have, until recently, covered the configurations that arise in
 1155 practice with sufficient frequency to warrant doctrinal investment.

1156 The configuration that AI-extensible products generate fits neither side cleanly. The multiple
 1157 users—eA through eD of Scenario γ , eA through eF of Scenario ε —do not share a common plan.
 1158 They do not share, in any meaningful sense, awareness of each other's contribution. They are not, in
 1159 that sense, intentional co-perpetrators. But they are not concurrent independent actors either: their
 1160 contributions are not similar (some are intentional, some negligent, some innocent), the harm can-
 1161 not be apportioned to any one of them on a defensible basis, and the model's processing makes the
 1162 contributions interact rather than merely concurring. The doctrine has no settled category for "simul-
 1163 taneous independent actors with different mental states whose conduct combines through an opaque
 1164 generative intermediary."

1165 This is the structural source of the mixed-culpability problem. It is not that the doctrine reaches
 1166 the wrong result; it is that the doctrine does not even reach a result, because the case sits outside its
 1167 operative categories.

1168 4.3.2 Three Inadequate Existing Tools

1169 Three doctrinal tools developed for adjacent problems are sometimes proposed as available for the
 1170 mixed-culpability case. Each addresses one dimension of the problem and falls short of addressing
 1171 the configuration as a whole.

1172 **Cumulative causation** (*kumulative Kausalität*) is the German doctrinal category for cases in
 1173 which two or more causal contributions, neither sufficient alone, combine to produce a harm. The
 1174 doctrine handles the causation dimension competently: each contributor's conduct is recognized as
 1175 a necessary part of the sufficient causal set, and liability can be predicated on each in turn. What the
 1176 doctrine does not address is the mental-state dimension: it operates within a single offense type and
 1177 presupposes that the constitutive elements (intent or negligence, as the offense requires) are present in
 1178 each contributor. It does not tell us how to handle the case in which one contributor acts intentionally,
 1179 another negligently, and a third innocently, with the offense available for some mental states but not
 1180 others.

1181 **Objective imputation** (*objektive Zurechnung*), the elaboration of post-causation normative filters
 1182 most associated with Roxin, addresses each actor separately rather than the relationship among them.
 1183 The doctrine asks, of each putative perpetrator, whether the conduct created or increased a relevant

risk that materialized in the result and whether the result lay within the protective purpose of the violated norm. Applied to Scenario γ , the doctrine produces defensible outcomes for each actor considered alone—eA and eD are intentional perpetrators (or fail to be so only on the common-plan dimension); eB’s case is the hard one of negligent contribution to an intentional offense; eC is exonerated. What it does not produce is an account of the actors’ relationship to one another, and in particular of how the joint operation of their contributions, mediated by the model, ought to be characterized for purposes of complicity doctrine.

Forensic exception by burden inversion is the most radical of the three. The Japanese Penal Code Art. 207 (同時傷害の特例) provides that where two or more persons inflict injury on a victim and the responsibility of each cannot be established, each shall be punished as if he had inflicted the injury alone. The provision solves the proof problem by inversion: the prosecution need not establish individual causation if it establishes the participation of each defendant in conduct of the relevant kind. The provision has been narrowly construed and is regarded by most commentators as a *sui generis* emergency provision rather than a model for general doctrine. Extending its logic to AI-extensible product contexts—punishing each contributor as if her contribution had solely produced the harm—would produce results that depart sharply from the proportionality and fault principles to which most modern criminal-law systems are committed. The article argues against this extension in Chapter 8.

None of the three tools, taken alone, suffices. The mixed-culpability case requires a doctrine that simultaneously handles aggregated causation, heterogeneous mental states, and the absence of common plan. The next subsection sketches what such a framework might look like, while emphasizing that the sketch is a framework for organizing analysis rather than a settled doctrinal category.

4.3.3 *Proposing a Layered Attribution Approach*

The article proposes a layered framework for organizing the mixed-culpability analysis without claiming to resolve it. The framework has four steps, each of which corresponds to a question that existing doctrine answers competently when posed in isolation but that current practice tends to conflate when posed in combination.

Step one: assess each agent’s mental state independently. The relevant mental states—intent, knowledge, recklessness or its civil-law analogues, negligence, innocence—are determined for each participant on the facts available to her, without reference to the other participants. eA’s intent is to defraud, eB’s mental state is at most negligent, eC’s is innocent, eD’s is intent shared with eA’s general purpose but not with the specific plan against V. This step prevents the analytical error of attributing one participant’s mental state to another simply because their conduct combined.

Step two: determine each agent’s objective causal contribution. The contribution is characterized in qualitative rather than quantitative terms, because the underlying processing is opaque to quantitative analysis. The relevant question is not “what percentage of the harm is attributable to this contribution?” but “did this contribution play a non-trivial role in the production of the harmful output?” This step accommodates black-box causation by abandoning the apportionment ambition and adopting a threshold contribution test. The threshold can be calibrated by reference to the norm violated: lower for offenses that aim at deterrence of risky conduct, higher for offenses that aim at

1224 attribution of fault for result.

1225 **Step three: apply the appropriate liability test for each agent separately.** Each participant is
 1226 then evaluated under the doctrine that her mental state and the offense at issue make applicable. eA is
 1227 evaluated under the indirect-perpetrator rules for intentional fraud. eB is evaluated under negligent-
 1228 tort doctrine in the tort claim and—if no negligent variant of the offense exists—excluded from crim-
 1229 inal liability. eC is excluded from both. eD is evaluated like eA, with the question of co-perpetration
 1230 with eA held over to step four.

1231 **Step four: address the relationship problem.** Where two or more participants are individually
 1232 liable on step three, the question of their relationship to one another remains. The framework dis-
 1233 tinguishes three categories: *co-principals* (where common plan or mutual awareness is established),
 1234 *accomplices* (where one's contribution facilitates the other's principal conduct), and *concurrent in-*
 1235 *dependent actors* (where neither relationship obtains). The mixed-culpability case typically falls into
 1236 the third category, with the doctrinal consequence that each is liable for her own contribution rather
 1237 than for the totality of the harm; the totality is recovered, if at all, through aggregating the individual
 1238 liabilities or through provider liability that captures the aggregation effect.

1239 This four-step framework is, again, not a doctrinal solution but a framework for organizing the
 1240 analysis. Its principal contribution is to separate questions that existing doctrine tends to conflate—
 1241 mental state, causal contribution, individual liability, relational liability—and to make explicit that
 1242 the relational question, in the AI-extensible context, will frequently be answered by “concurrent inde-
 1243 pendent actors” rather than by either of the more familiar categories. The implications of that finding
 1244 for substantive doctrine and for policy are taken up in Chapters 5 and 8.

1245 4.4 Cross-Agent Liability Allocation

1246 4.4.1 Three Classes of Agent, Three Tools, No Framework

1247 Every harm generated by an AI-extensible product implicates three classes of agent. The first is the
 1248 *provider*—the legal entity, typically a corporation, that designs, deploys, and operates the product.
 1249 Company B of the running scenarios is this party. The provider is, in the doctrinal sense, the most
 1250 clearly attributable of the three: it is incorporated, has assets, and operates within the regulatory
 1251 reach of identifiable states. The second class is the *user or users*—natural persons (or, in business-
 1252 to-business deployments, employee-agents of legal persons) who submit prompts. The user class is
 1253 heterogeneous in mental state, geographic location, and reachability. The third class is the *AI system*
 1254 *itself*. This class lacks legal personhood in every jurisdiction surveyed; the *Thaler v. Perlmutter* line
 1255 of US cases, the German Federal Patent Court's DABUS decisions, and the EU's deliberate refusal in
 1256 the AI Act to confer “electronic personhood” all confirm the position. Liability theory must therefore
 1257 distribute the responsibility the AI system might otherwise bear to the other two classes (or, in some
 1258 accounts, allow it to fall to the victim).

1259 The doctrinal apparatus has produced a number of tools for allocating among these classes, none
 1260 of which integrates the others. The *intermediary safe-harbor model*—DMCA §512 in the United
 1261 States, the EU E-Commerce Directive (now succeeded for many purposes by the Digital Services
 1262 Act), Japan's プロバイダ責任制限法—was designed for passive hosting of user-generated content

and conditions immunity on knowledge-based takedown. The *neutral-tool provider doctrine*, exemplified by the Japanese Supreme Court's *Winny* judgment of 2011, allocates responsibility based on the provider's specific subjective awareness that the tool would be used predominantly for unlawful purposes. The *risk-graded ex-ante regulation model*, exemplified by the EU AI Act of 2024, imposes obligations calibrated to the use-case risk level rather than to particular incidents. The *common-law contributory and vicarious liability frameworks*, developed in IP and tort, allocate based on contribution and control.

Each tool was developed for a different paradigm and addresses a different aspect of provider responsibility. None addresses the particular difficulty that AI-extensible products present: that the provider's contribution is neither passive hosting nor specific unlawful design, but the operation of a non-deterministic generative process whose outputs are co-shaped by users whose own attribution is itself contested. The tools are partial answers to a problem that their designers did not foresee and cannot, on their own terms, fully engage.

4.4.2 The Insufficiency of Current Tools

The mismatch between the available tools and the AI-extensible product paradigm is not a contingent feature of poorly drafted statutes; it follows from the architecture of each tool.

The intermediary safe-harbor regimes presuppose a passive role for the intermediary: the platform hosts content created by users, and its liability is conditioned on actual or constructive knowledge of the content's unlawfulness. This conception fits the YouTube or Wikipedia case well; it fits the AI-extensible product case poorly. The provider of an AI-extensible product is not passively hosting user-supplied output; it is generating output through a process that the provider designed, trained, and operates. The output is, in a real sense, the provider's product, even though it is co-shaped by user prompts. The doctrinal extension of safe-harbor immunity to this configuration—now contested in the United States with respect to Section 230's application to generative AI, and in the EU with respect to the DSA's hosting-service definition—requires either expanding the immunity beyond its conceptual basis or limiting it in ways that lose its operational simplicity.

The neutral-tool provider doctrine, in its *Winny* formulation, conditions provider liability on specific subjective awareness that the tool was provided with the principal purpose of unlawful use. The doctrine fits a paradigm in which the tool's behavior is deterministic and the unlawful use is foreseeable as such. It fits poorly a paradigm in which the tool's behavior is emergent and unpredictable per particular interaction, even though the aggregate distribution of harmful and lawful uses is statistically tractable. A provider of a general-purpose AI assistant cannot point to a specific subjective intention to facilitate harm; nor can prosecutors prove that the provider's principal purpose was unlawful use. The *Winny* doctrine, on its face, exonerates such providers; the question whether such exoneration is adequate to victim protection is the policy question Chapter 8 engages.

The EU AI Act's risk-graded approach was designed to address ex-ante risk through compliance obligations rather than ex-post attribution through liability rules. The Act requires high-risk-system providers to implement risk-management procedures, technical documentation, data-governance measures, transparency, and post-market monitoring; it does not allocate liability when these procedures, properly implemented, nonetheless yield harmful outputs. The companion AI Liability Directive

proposal addresses the ex-post side, principally through disclosure obligations and rebuttable causal presumptions, but the proposal has had a difficult legislative path and its final form is uncertain. Even in its full proposed form, the AILD would address provider-defendant cases without providing a framework for the multi-user attribution question that lies at the heart of the mixed-culpability problem.

The result is that current tools are partial answers, each addressing one face of the cross-agent allocation problem and none integrating the others. The article does not propose a unified replacement; the proposal would require a degree of cross-doctrinal harmonization that no jurisdiction has yet undertaken. It does propose, in Chapter 8, a set of principles that could guide incremental convergence.

4.4.3 Connecting to the Case Studies

The cross-agent allocation problem interacts distinctively with the other two structural sources in each of the constructed scenarios.

In Scenario α , cross-agent allocation is at its simplest: eA's clear primary liability allows the provider to invoke the absence of specific knowledge as a basis for non-liability. Black-box causation and mixed culpability are dormant. The configuration is the easy case of provider non-attribution.

In Scenario β , the user's negligence creates a wider field for provider co-attribution. If the negligent contribution flowed in part from the provider's failure to constrain the prompt-engineering surface in ways a reasonable design would have constrained it, contributory liability is plausible. Black-box causation operates only weakly here, because eB's contribution is a single user's prompt that the provider's own systems can in principle examine.

In Scenario γ , the multi-user mixed-culpability case exposes the cross-agent allocation problem most acutely. The provider is, on the facts, the only actor who could have foreseen and prevented the aggregation effect: the model's shared retrieval-augmented context is a design choice, and the prospect that prompts from different users could combine to produce harmful outputs is a foreseeable consequence of the design. The doctrinal question is whether the foreseeability of the *class* of aggregation harms suffices for liability when the particular harm in V's case was, by hypothesis, not specifically foreseeable. Black-box causation amplifies the difficulty: even if the provider's design choice is shown to have been negligent, the contribution of the design choice to V's harm cannot be quantified. Mixed culpability adds the dimension of allocating among the provider and the user-actors who are also liable.

In Scenario δ and Scenario ε , the cross-agent allocation problem combines with cross-jurisdictional dispersion. The provider may face liability in some jurisdictions and not others, on bases that vary with each forum's calibration of provider responsibility. The result, traced in detail in §3.6, is that the provider's exposure is heavily a function of which forum prevails—a function that is itself contingent on enforcement realities orthogonal to the merits.

The cross-agent allocation problem is, in this sense, the structural source most sensitive to the configuration of the other two. Its analysis cannot be conducted in isolation from black-box causation and mixed culpability; the three sources operate jointly, and the doctrinal apparatus must engage them as a system rather than serially.

4.5 The Homology Across Doctrines: A Synthesis

The three structural sources are not a closed list. A more extensive analysis would identify additional sources—perhaps the temporal asymmetry between the speed of model output and the slowness of legal adjudication, perhaps the institutional asymmetry between provider expertise and judicial expertise, perhaps the political-economy asymmetry between provider lobbying capacity and victim advocacy. The three sources here selected, however, capture the principal mechanisms by which existing doctrine fails on the configurations the article analyzes; they are also the sources whose homology across multiple doctrines is most pronounced. The article therefore organizes its doctrinal analysis around them.

The next chapter turns from the cross-doctrinal generality of these sources to their doctrine-specific manifestations. Each of the seven fields the article surveys engages each structural source in its own terms: criminal law confronts black-box causation through the standards of proof and the requirements of the constitutive offense; tort confronts mixed culpability through joint-tortfeasance doctrine; copyright confronts cross-agent allocation through the joint-authorship test; and so on. The doctrine-by-doctrine analysis of Chapter 5 maps these manifestations field by field; the comparative analysis of Chapter 6 then traces how different jurisdictions have begun to engage them.

5. Doctrinal Dilemmas in Specific Legal Fields

Introductory Note

This chapter complements the cross-doctrinal analysis of Chapter 4 by working through each of the seven doctrinal fields the article surveys, asking how each engages the three structural sources in its own terms. The treatment is asymmetric. Criminal law, tort law, and intellectual property receive in-depth treatment because (a) their dilemmas have received the most sustained scholarly attention in adjacent domains; (b) they generate the most distinctive AI-extensible-product problems; and (c) they form the core of the comparative analysis in Chapter 6. Contract law, data protection, administrative regulation, and private international law receive briefer treatment, sufficient to map the principal difficulties without rehearsing what specialist literatures already treat at length. Two further fields—labor law and consumer protection law—receive only brief mention as flagged open questions.

5.1 Criminal Law

5.1.1 The Indirect Perpetrator Framework and Its Limits

The doctrine of indirect perpetration—*mittelbare Täterschaft* in German law, 間接正犯 in Japanese, indirect perpetration via innocent or *doli incapax* agent in common-law systems, and the analogous principal-by-agent doctrines in US federal criminal law—was developed for cases in which a perpetrator uses another person, who lacks full criminal capacity for the relevant offense, as a tool to accomplish her unlawful purpose. The classic case is the unwitting messenger who delivers a poisoned package; the modern elaborations cover deception of a person who lacks knowledge of the offending circumstance, exploitation of a person under duress, and direction of a person within an

organized power apparatus that ensures execution.

Mapping the doctrine onto the AI-extensible product scenario is straightforward in its central application. eA of Scenario α uses the AI system as a tool—more precisely, as an instrumentality through which she conveys her deceptive communication—and her liability is grounded in indirect perpetration on a strict tool theory (the AI is not a person but a mechanism, like a printing press or a firearm) or in the constituent fraud offense itself, with the AI as the medium of the deception. Either route reaches the right result on Scenario α 's facts. The doctrine handles single-user intentional misuse with little strain.

The doctrine breaks down in three configurations the article identifies. First, where multiple users contribute prompts that interact through the model's processing, the indirect-perpetrator analysis cannot be applied to each in turn without producing the implausible result that each is solely responsible for the entire output. The doctrine presupposes a perpetrator who uses the tool; it does not accommodate multiple perpetrators using the same tool simultaneously, each of whose contributions shapes the result. Second, where the prompt is ambiguous and the model's processing introduces content the prompter did not specifically intend, the line between perpetrator-direction and tool-emergence becomes unstable; the doctrine has not yet worked out which side of the line the AI-generated portion falls on. Third, where the user's mental state is negligent rather than intentional, indirect perpetration is unavailable (it is a doctrine of intentional liability), and the analysis must shift to negligent-result offenses that may or may not exist for the relevant constituent offense.

The Roxin elaboration of indirect perpetration via *Organisationsherrschaft*—perpetration through an organized apparatus of power—offers an interesting comparative reference point. Roxin developed the doctrine for cases such as the *Mauerschützen* (border guards' shooting at the Berlin Wall inner-German border) where the immediate actor is criminally responsible but the perpetrator's control is exercised through an organizational structure that ensures the act will be performed. The analogy to a provider whose product reliably produces a class of outputs has been suggested in the AI literature; the analogy is suggestive but limited, because the provider's control over particular outputs is statistical rather than reliable. The doctrine illuminates what is missing in the AI-extensible product context—reliable control of the result through the intermediary—more than it provides a tool for handling cases where reliable control is absent.

5.1.2 Mixed Culpability in Penal Code Doctrine

The major civil-law penal codes engage the mixed-mental-state scenario through the doctrines of co-perpetration, accomplice liability, and neutral-tool provision. None of the three engages the configuration of Scenario γ cleanly.

In Japanese criminal-law theory, 共同正犯 (co-perpetration, Penal Code Art. 60) requires 意思の連絡 (mutual awareness) between the participants. eA and eD of Scenario γ share a general fraudulent purpose but lack mutual awareness of each other's specific contributions; the leading commentators (松宮孝明, 佐伯仁志, 山口厚) agree that the requirement is not satisfied by mere parallel intent. The category of 同時犯 (concurrent offender) handles cases of independent action without mutual awareness, but it does not address how the contributions of multiple concurrent offenders should be attributed when their conduct interacts through an intermediary process. eB's negligent contribution

falls outside both categories because there is no negligent variant of fraud (過失詐欺罪 does not exist).

The *Winny* judgment of the Japanese Supreme Court (最決平成 23 年 12 月 19 日) supplies the principal Japanese provider-liability standard. The Court held that the developer of the Winny file-sharing software could be convicted of complicity in copyright infringement only if he provided the software with the specific subjective awareness that it would be used predominantly for infringing purposes; mere awareness of substantial possible misuse was insufficient. The standard is highly favorable to providers, and on the facts of Scenario γ would clearly exonerate Company B. The doctrinal cost of this exoneration—in light of the cumulative harm across users—is one of the central tensions the article identifies.

In German criminal-law theory, *Mittäterschaft* (§25(2) StGB) requires *gemeinsamer Tatentschluss* (common plan) and *gemeinsame Tatbegehung* (joint execution). The BGH jurisprudence on the requirement of common plan is exacting: parallel intent without mutual awareness does not suffice (*BGHSt* 32, 25; *BGHSt* 49, 39). *Anstiftung* (§26) and *Beihilfe* (§27) require subjective awareness of the principal offense; eB's negligence does not satisfy either. *Nebentäterschaft* (concurrent perpetration) handles independent contributions but, as in Japanese law, does not specify how to handle interactive contributions mediated by an opaque process. *Fahrlässige Mittäterschaft* (negligent co-perpetration) has been the subject of extended doctrinal debate (Roxin, Kindhäuser, Renzikowski) but remains unsettled and, in any case, requires the underlying offense to admit a negligent variant.

In Chinese criminal-law doctrine, the absence of a generally applicable common-law-style aiding-and-abetting category is supplied by Art. 25–29 on 共同犯罪 (joint crime), which generally requires intentional joint action, and by the network-specific offense of 帮助信息网络犯罪活动罪 (Art. 287-2, “帮信罪”), introduced in the 2015 amendment to the Criminal Law and clarified by the Two Highs' 2019 Judicial Interpretation. 帮信罪 criminalizes the knowing provision of network technical support, advertising promotion, payment settlement, or similar assistance to others' information network crimes. The 2019 Interpretation lowers the threshold for 明知 (knowledge) by permitting inference from objective circumstances, including transaction patterns, repeat dealings, and abnormal compensation arrangements. The offense potentially captures provider conduct in AI-extensible product harm cases that the *Winny* doctrine in Japan would exclude; but its scope and the inference rules for knowledge remain contested in Chinese scholarship (张明楷, 陈兴良). The article returns to this divergence in Chapter 6.

5.1.3 The Causation Burden Under Black-Box Conditions

The standard of causal proof in criminal proceedings is everywhere high. The Anglo-American formulation—“proof beyond reasonable doubt”—is matched in the civil-law systems by formulations such as the German *volle richterliche Überzeugung* (full judicial conviction), the Japanese 合理的な疑いを差し挟まない程度の証明 (proof that does not admit reasonable doubt), and the Chinese 排除合理怀疑 (excluding reasonable doubt) standard formalized in Criminal Procedure Law Art. 55. The standard applies to each constitutive element of the offense, including—where the offense requires it—the causal contribution of the defendant's conduct to the prohibited result.

Black-box causation presents a configuration in which this standard is not merely difficult to sat-

isfy but, on the analysis of §4.2, structurally unsatisfiable in important classes of case. Where multiple users have contributed prompts that have been aggregated through an opaque generative process to produce the harmful output, the share of the output attributable to any one user's contribution is not estimable to the standard of confidence the criminal-law standard requires. The result is that the prosecution that bears the burden of proof simply cannot meet it, even where the defendant's conduct is morally blameworthy and his contribution to the harm is, in some weaker sense, established.

The temptation, in such a configuration, is to relax the standard of proof or to invert the burden. Several proposals along these lines have been advanced in the AI-law literature: importing Japanese-Penal-Code-Art. 207-style burden inversion to the AI context, adopting EU-AI-Liability-Directive-style rebuttable causal presumptions for criminal as well as civil contexts, recognizing constructive causation through participation in a course of conduct likely to produce the harm. The article argues against each of these extensions to criminal law, and the argument has the same shape in each case: the criminal-law standards of proof are not technical inconveniences to be relaxed when they are inconvenient; they are substantive commitments of criminal law to the protection of the innocent and to the proportionality of punishment. The configurations in which they cannot be satisfied are configurations in which criminal liability ought not to attach. The proper response is not to lower the bar but to route the case through civil and administrative channels that do not require the same standard.

This is the meaning, on the article's argument, of criminal law's commitment to *ultima ratio*. The commitment is not merely that criminal law should be the last resort to be invoked, but that where its evidentiary preconditions cannot be satisfied, criminal liability should fail and other remedies should bear the weight. The black-box causation problem in the AI-extensible product context is, on this view, not a defect of criminal law to be remedied but a constraint that criminal law's commitment to humility should respect. The policy implications of this position are developed in Chapter 8.

5.2 Tort Liability

5.2.1 Negligence, Foreseeability, and Joint Tortfeasance

The general tort doctrines that bear on AI-extensible product harm share a common conceptual framework even where their statutory expression varies. The Japanese Civil Code Art. 709 imposes liability on one who, intentionally or negligently, infringes another's right or legally protected interest and causes damage. §823 BGB imposes liability for intentional or negligent injury to enumerated protected interests. The Anglo-American common law of negligence requires duty, breach, causation, and damage; the Restatement (Third) of Torts codifies the analysis in updated form. Each framework requires (i) a duty owed by the defendant to the plaintiff, (ii) breach of that duty through conduct falling below the applicable standard of care, (iii) causation between breach and harm, and (iv) cognizable harm.

The foreseeability standard operates at two points in the negligence analysis: in the determination of duty (whether the harm to the plaintiff was within the range of harms a reasonable person in the defendant's position should have anticipated) and in the determination of proximate cause (whether the harm that materialized was sufficiently connected to the breach to be legally attributed to it).

The AI-extensible product context strains the foreseeability inquiry distinctively. The aggregation effect by which multiple users' prompts combine to produce harmful outputs is, at the class level, foreseeable to a sophisticated provider; at the level of the particular harm to the particular victim, it is not. The professional security researcher of Scenario β can foresee that publication of effective persuasion techniques may aid social-engineering fraudsters; she cannot foresee any specific victim's loss.

The duty-of-care standard is itself contested where the actor is a provider of a generative-AI system. The Hand-formula tradition in US tort law calibrates duty by reference to the burden of avoidance, the probability of harm, and the magnitude of harm if it occurs. Applied to AI providers, the formula's denominator is large (the burden of avoiding all possible harms through restrictive product design is high) but its numerator may also be substantial (the probability of harm across a large user base, summed over many particular harms, can be quantified). The civil-law duty-of-care analysis, more often articulated in terms of the *Verkehrspflichten* (traffic obligations) of those who create sources of danger, similarly imposes obligations calibrated to the foreseeable risks of the activity. Neither tradition has yet produced a settled framework for AI-provider duty.

The joint-tortfeasance analysis adds further complications, taken up in the next subsection.

5.2.2 The Joint Tortfeasance Problem in Scenario Gamma

The treatment of multiple tortfeasors varies across civil-law and common-law systems but converges on a common difficulty when applied to Scenario γ 's mixed-culpability configuration.

The German doctrine distinguishes *Mittäterschaft* of §830(1) sentence 1 BGB (joint tort by common plan), *Beteiligung* of §830(2) (instigators and accomplices, treated as co-tortfeasors), and *Nebentäterschaft* (concurrent independent tortfeasors, each liable for the harm she individually caused). The first two require subjective coordination among the tortfeasors that Scenario γ 's participants lack; the third results in each contributor being liable only for her own causal contribution, which under black-box conditions cannot be quantified. Markesinis and Unberath's comparative treatment notes that the German courts have responded to similar difficulties in other contexts (mass-tort and environmental harm) through evidentiary devices that shift the burden of proof to the defendant; the analogous extension to the AI context is contested.

The Japanese doctrine under Civil Code Art. 719 distinguishes 共同不法行為 (joint tort) and the closely related configurations of joint causation and uncertain attribution. Art. 719 para. 1 first sentence requires 共同 (jointness, understood in the dominant view as objective relatedness rather than subjective coordination); para. 1 second sentence handles cases where the specific contributor cannot be identified by holding all participants jointly liable. The doctrine reaches further than its German analogue in some respects but leaves the same gap: it does not address how contributions of different mental qualities should be aggregated when the mediating process is opaque.

The Anglo-American doctrine of joint and several liability, modified in many US jurisdictions by comparative-fault apportionment statutes, permits the plaintiff to recover the entirety of damages from any liable defendant, with the defendants left to apportion among themselves. The doctrine fits the AI context awkwardly because the apportionment cannot be defensibly performed: the contributor whose prompt was, on a defensible measurement, twenty percent responsible ends up bearing one-hundred-

percent liability under joint-and-several rules in the absence of a quantitative basis for shifting the rest.

The 2024 revision of the EU Product Liability Directive (Directive (EU) 2024/2853, repealing the 1985 Directive) extends strict product-liability principles to software and AI-enabled products; this is a substantive expansion that materially alters the analysis on the provider side. The accompanying proposed AI Liability Directive, which would have introduced rebuttable causal presumptions in fault-based AI-related claims, has had a difficult legislative trajectory and at the time of writing remains unsettled. If adopted, the AILD would meaningfully shift the tort analysis in provider-defendant cases by lowering the plaintiff's evidentiary burden in respect of causation; it would not, however, resolve the multi-user joint-tortfeasance configuration that is this article's principal concern.

5.2.3 Causation in Civil Litigation

The civil standard of proof (preponderance of the evidence in common-law systems, *überwiegende Wahrscheinlichkeit* or its equivalent in civil-law systems) is, on its face, more accommodating of imperfect causal knowledge than the criminal standard. The plaintiff need only show that more likely than not the defendant's conduct caused the harm; the gap between certainty and proven likelihood is, in most cases, narrower than the gap that the criminal standard requires the prosecution to bridge.

But the standard is, in the black-box-causation configuration, analogous to the criminal standard in its underlying difficulty: it requires the plaintiff to make defensible quantitative claims about causal contribution that the underlying technology does not support. The plaintiff cannot show that the defendant's conduct more likely than not caused the harm if the share of the harm attributable to the defendant's contribution is constitutively unascertainable. The lower standard does not rescue the analysis when the rescue would require quantitative information the technology does not yield.

Several adaptations developed for analogous configurations in the mass-tort context are sometimes proposed as analogs. *Sindell v. Abbott Laboratories* (Cal. 1980) introduced market-share liability for cases in which DES-related injuries could be traced to the drug class but not to the particular manufacturer; each manufacturer was held liable in proportion to its market share of the relevant DES generation. The reasoning has been applied—haltingly—in adjacent contexts and rejected in others. The transfer to the AI-extensible product context is non-obvious. The analog of the “manufacturer” would be the user-prompt source; the analog of “market share” would be some defensible quantification of contribution to the output. But the second analog fails: there is no defensible quantification.

The loss-of-chance doctrine, developed in the *Herskovits v. Group Health Cooperative* line of US medical malpractice cases and the comparable European decisions on *verlorene Heilungschance*, permits recovery for the lost chance of a better outcome where the defendant's conduct reduced the chance of recovery. The analogy to AI cases would require treating the user's contribution as having altered the probability of a particular harmful output; the analogy is suggestive but again requires quantitative information the underlying processing does not yield.

The most promising adaptation is the EU AILD's proposed rebuttable causal presumption: where the plaintiff shows fault on the part of the AI provider and a connection between that fault and the kind of harm suffered, causation is presumed unless the provider rebuts. The device shifts the evidentiary burden without claiming to ascertain the actually unascertainable. Whether the device should

extend beyond the provider-defendant context to the multi-user joint-tortfeasance configuration—and whether civil-law jurisdictions outside the EU will adopt analogous mechanisms—are questions Chapter 6 engages.

5.3 Intellectual Property: Copyright and Beyond

5.3.1 The Authorship Question

The copyright analysis of AI-extensible product output begins with the authorship question: whether copyright subsists in the output at all, and if so, in whom. The question implicates the personhood predicate of authorship, the contribution analysis of joint authorship, and the treatment of derivative works.

The personhood predicate is the most settled. Across all six jurisdictions surveyed, copyright presupposes a human (or, in some formulations, at least non-machine) author. In the United States, *Thaler v. Perlmutter* (DDC 2023, aff'd DC Cir. 2025) confirmed that the Copyright Office's long-standing refusal to register AI-generated works without human authorship is consistent with the Copyright Act; the Office's three-installment guidance of 2023 elaborates the application of this position to mixed human-AI works through the *Zarya of the Dawn* registration decision, which allowed registration of a graphic-novel arrangement and text but denied it for the AI-generated images. In Japan, the Copyright Act's Art. 2(1)(i) defines a copyrightable work as "a creative production of thought or sentiment," a formulation read by the Cultural Affairs Agency's 2024 *AI と著作権に関する考え方* *nitsuite* as requiring human creative intervention; AI-generated portions without human creative direction are not copyrightable. In the European Union, the CJEU's *Painer* and *Infopaq* jurisprudence on "author's own intellectual creation" has been read in member-state implementations as requiring a human creator. China's Copyright Law Art. 3 defines copyrightable works in terms that have been read more flexibly by lower courts; the Beijing Internet Court's 2023 judgment in *Li v. Liu* recognized copyright in an AI-generated image where the human user's contribution—through detailed prompts, parameter selection, and post-generation modification—was substantial and creative.

The contribution analysis is less settled. Where multiple users contribute prompts that, mediated by the model, produce a single output, the joint-authorship test of US copyright (Childress test: intent to merge contributions and independently copyrightable contributions from each putative author) and its civil-law analogues do not produce determinate results. The first prong fails because the users typically lack subjective awareness of one another. The second prong is contested: a single prompt may or may not be independently copyrightable depending on its specificity, and the black-box causation problem prevents determinate measurement of which prompts contributed which expressive elements to the output.

The derivative-works framework—Japan Copyright Act Art. 27–28 on translation, adaptation, and derivative-work rights; US 17 USC §103 on protection of derivative works to the extent of new authorship; China Copyright Law Art. 13 on adaptation rights; EU InfoSoc Directive Art. 2–4 on the reproduction and communication rights—supplies a familiar conceptual apparatus but does not by itself resolve the AI-output case. The output is, in some sense, a derivative work of the underlying training data and of the user prompts; but the derivative-works doctrine presupposes identifiable

underlying works and identifiable derivative authorship, neither of which is reliably present.

The result is a doctrinal landscape in which the central question —whether copyright subsists in AI-extensible product output and in whom—is answered differently in each jurisdiction surveyed and is in flux even within each. The article does not propose a uniform answer; it argues that the question is itself a manifestation of the cross-agent allocation problem in copyright form, and that any answer will need to cohere with the broader allocation framework adopted in the jurisdiction.

5.3.2 Three-Layer Copyright in AI-Extensible Outputs

The article proposes a three-layer model for the copyright analysis of AI-extensible product outputs. The model is descriptive rather than prescriptive: it organizes the layers of authorship and rights that contemporary doctrine recognizes (or fails to recognize) in the output, and surfaces the doctrinal questions each layer raises.

The first layer is the *provider's copyright in the underlying software, model architecture, and product surface*. This layer is uncontroversial. Company B's copyright in its codebase, its fine-tuning configurations, and the structural elements of its product is unaffected by user activity. The provider's contribution to the output, however, is mediated by the model in ways that do not, on the prevailing copyright analyses, give the provider a separate copyright in the output: the model's contribution is autonomous in the relevant sense, and the provider's authorship in the model is not extended to the model's products.

The second layer is the *user's copyright in any creative contribution she made through her prompts*, where recognized. This layer's recognition varies by jurisdiction (the Beijing Internet Court's 2023 judgment recognizes it generously where the user's contribution is substantial; the US Copyright Office's 2023 guidance recognizes it more narrowly). Where multiple users contribute, the attribution problem of the previous subsection makes layer-two analysis indeterminate even within jurisdictions that recognize prompt-based authorship.

The third layer is the *AI-autonomous portion of the output*, which on the prevailing analyses falls into a no-rights zone: it is not the provider's because the provider did not author it directly; it is not the user's because the user did not author it; it is not the AI's because the AI lacks legal personhood. The no-rights zone has both positive and negative consequences. Positively, the output falls into the public domain to the extent of the AI-autonomous portion, with the corresponding copying freedoms. Negatively, no party has standing to enforce against unauthorized use, and the user who relies on the output as proprietary discovers that her reliance is unfounded.

The layered structure interacts with the ToS architecture of the product in complex ways. Most provider ToS attempt to assign all intellectual property rights in user inputs and outputs to the provider, sometimes via license-back provisions to the user. These provisions can effectively allocate layer-two rights between provider and user; they cannot effectively allocate layer-three rights because no party holds them in the first place. The result is a contractually constructed scheme of rights that overlies a doctrinal landscape in which the underlying rights are partly indeterminate and partly non-existent. The discrepancy is a source of latent litigation risk that has, at the time of writing, only begun to surface in actual cases.

5.3.3 Beyond Copyright: Trade Secrets, Patents, Trademarks

The non-copyright branches of intellectual property are affected by AI-extensible products to a more limited extent.

Patent law is largely unaffected on the inventorship side. The DABUS line of cases—decided uniformly against AI inventorship in the United States, the United Kingdom, the EU through the EPO, and Australia (with a brief reversal that was itself reversed)—confirms that the “inventor” on a patent application must be a natural person. The use of AI as a research aid does not, on the prevailing analyses, defeat patentability so long as a human inventor’s contribution to the inventive concept is identifiable. The principal patent-law question raised by AI-extensible products is the more mundane question of disclosure: whether a patent applicant who used an AI tool in formulating the inventive concept must so disclose, and how the obviousness analysis is affected by widely available AI tools. These questions remain unsettled.

Trade-secret law is implicated wherever prompts to AI-extensible products contain confidential information. The transmission of trade secrets to a generative-AI service may waive the secrecy required for trade-secret protection, depending on the service’s confidentiality posture; the service’s possible use of the input for training may further compromise the secret. The recent enterprise-grade offerings of major AI providers explicitly disclaim training on customer inputs, which addresses one face of the problem; the other face—the risk of inadvertent disclosure through model outputs—remains.

Trademark law is implicated when AI-generated outputs use protected marks. The doctrine of likelihood of confusion does not distinguish between human-authored and AI-authored uses; the user who prompts an AI to produce a mark-bearing output bears the trademark liability she would bear had she produced the output herself. The provider’s secondary liability tracks the contributory and vicarious infringement frameworks that apply to other intermediation configurations.

In each of these branches, the analysis is doctrinally more straightforward than in copyright, principally because the authorship-personhood question that destabilizes copyright analysis does not arise in the same form. The cross-agent allocation problem nonetheless recurs, in particular for provider secondary liability under each branch.

5.4 Contract Law and ToS Architecture

Provider Terms of Service and end-user licence agreements attempt to allocate the risks AI-extensible products generate. Typical provisions include user-generated-content disclaimers (the user warrants that her prompts do not infringe third-party rights), indemnification clauses (the user indemnifies the provider for liability arising from her use), intellectual-property assignment or licence-back arrangements addressed in §5.3, prohibited-use catalogues (fraud, harassment, child-sexual-abuse material, etc.), and liability-limitation provisions that cap the provider’s exposure at, typically, fees paid in the preceding twelve months.

The enforceability of these provisions is constrained by the unfair-contract-terms regimes that apply in each jurisdiction. German *AGB-Recht* (BGB §§305–310) treats provider-drafted clauses as standard terms subject to substantive content control; clauses that unreasonably disadvantage the

user are invalid. Japanese 消費者契約法 similarly invalidates clauses that unreasonably restrict consumer rights; the more recent 民法 amendments on standard terms (Art. 548-2 to 548-4) extend related controls. The EU Unfair Contract Terms Directive 93/13/EEC and the Digital Content Directive 2019/770 impose analogous controls on B2C providers across the EU. China's 消费者权益保护法 and Civil Code Art. 496–498 perform the analogous function. The US is the principal exception: contract freedom is broader, and clauses that would be invalid in Europe or Japan often survive scrutiny in US courts.

The substantive question this article foregrounds is more specific: can a ToS effectively allocate criminal or tort liability to users in the AI-extensible product context? The answer, broadly, is that ToS provisions can effectively allocate *contractual* liability and can effectively allocate *intellectual-property* entitlements, but cannot effectively allocate liability for harms suffered by third parties (V in our running scenarios) who are not parties to the ToS. Third-party harm liability is governed by the underlying tort or criminal law and is not displaceable by contract between the immediate parties; this is a basic limit of contractual allocation. Provider indemnification clauses can shift the ultimate financial bearing of that liability among provider and user, but cannot extinguish the underlying liability to the third party.

5.5 Data Protection and Personal Information

The data-protection regimes implicated by AI-extensible products operate on overlapping but non-identical conceptual foundations. The EU GDPR (Regulation 2016/679) imposes obligations on data controllers and processors with respect to “personal data,” broadly defined. The Japanese 個人情報保護法 (APPI), most recently amended in 2022 and 2023, parallels the GDPR in many respects but with distinctive provisions on “personal information that requires particular care.” China's PIPL of 2021 imposes consent and purpose limitations broadly comparable to the GDPR, with stricter cross-border-transfer controls. The United States lacks a comprehensive federal data-protection statute and proceeds through sectoral regimes (HIPAA for health, GLBA for financial, COPPA for children's data) and an emerging patchwork of state laws (CCPA/CPRA, Virginia CDPA, etc.).

AI-extensible products pose four distinctive data-protection difficulties. *First*, user prompts frequently contain personal information about third parties who have not consented to processing: eA's prompts to draft messages to V contain V's identifying information; eB's professional security-research prompts may contain examples drawn from real victims. The processing of this information by the provider may lack a lawful basis under GDPR Art. 6 or its analogues. *Second*, AI outputs may inadvertently expose personal information from training data through the well-documented memorization phenomenon. The data-subject's remedies for such exposure are constrained by the difficulty of proving that the output reflects the data subject's training data rather than a coincidental generation. *Third*, multiple users' prompts may aggregate to identify a third party who is identifiable from no single prompt alone; the aggregation effect that characterizes Scenario γ replicates in the data-protection register. *Fourth*, the right to explanation in automated decision-making under GDPR Art. 22, and the analogous transparency obligations under PIPL and APPI, are unsatisfiable in the strict sense for black-box generative AI; the data controller cannot supply an explanation it does not itself possess. The doctrinal accommodations under development—explanation-as-purpose, explanation-as-

1734 process, explanation-as-counterfactual—each lose something the original right was meant to secure.

1735 5.6 Administrative Regulation

1736 The administrative regulation of AI-extensible products has, in the past three years, become the most
 1737 rapidly developing area of AI law. Five regimes warrant note. The EU AI Act (Regulation 2024/1689)
 1738 adopts a risk-tier approach with four levels (unacceptable, high, limited, minimal), differentiated
 1739 obligations for providers, deployers, importers, and distributors, and a penalty regime permitting fines
 1740 of up to seven percent of global annual turnover for the most serious violations. China’s 生成式人
 1741 工智能服务管理暂行办法 (Generative AI Service Management Interim Measures, 2023) imposes
 1742 consent, content-moderation, data-quality, and registration obligations on providers of generative-AI
 1743 services to the Chinese public. The UK Online Safety Act 2023 imposes duties of care on online
 1744 services and includes specific provisions reaching some AI-generated content; the contemporaneous
 1745 AI Regulation White Paper (2023) declines comprehensive horizontal regulation in favor of regulator-
 1746 led sectoral approaches. The US relies principally on the NIST AI Risk Management Framework
 1747 (2023, with the Generative AI Profile of 2024), sector-specific agency action, and emerging state
 1748 legislation (California, Colorado, Texas). Japan operates through the soft-law AI 事業者ガイド
 1749 ライン of the Ministry of Economy, Trade, and Industry and the Ministry of Internal Affairs and
 1750 Communications (2024).

1751 A common conceptual question across these regimes is the identification of the regulatory subject.
 1752 The EU AI Act’s most consequential choice—to distinguish “provider” from “deployer” and impose
 1753 risk-management obligations on both—is the most explicit attempt to allocate among the actors impli-
 1754 cated in an AI-extensible product deployment. The Act’s provider-deployer distinction maps imper-
 1755 fectly onto the AI-extensible product paradigm because the user, in the sense of the article’s running
 1756 scenarios, is typically not a “deployer” in the Act’s sense (deploying for professional purposes) but
 1757 an end-user whose interaction with the product is a single session. The user’s responsibilities under
 1758 the Act are correspondingly limited; the principal regulatory weight falls on the provider.

1759 The other regimes make analogous but not identical choices. China’s generative-AI rules fo-
 1760 cus principally on the provider; the UK OSA imposes duties of care on the operator of the online
 1761 service, with “operator” covering both providers and deployers in relevant respects; the US sectoral
 1762 approaches typically fix on the entity that implements AI in the regulated sector. The result is a multi-
 1763 axis regulatory landscape in which the same provider may face different obligations as “provider,”
 1764 “operator,” or “service” under overlapping regimes. Scenario ε ’s analysis traces the consequences
 1765 for the multinational provider configuration.

1766 5.7 Private International Law

1767 The cross-border dimension developed in Scenario δ and Scenario ε engages private international
 1768 law at three points: jurisdiction, applicable law, and recognition and enforcement of judgments.

1769 *Jurisdiction* in cases involving AI-extensible products is contested between the traditional terri-
 1770 torial bases (situs of conduct, situs of harm, defendant’s domicile) and bases adapted to ubiquitous
 1771 online services. The Brussels Ibis Regulation, the Lugano Convention, the Hague Convention on

Choice of Court Agreements (2005), and the domestic-law analogues in non-EU jurisdictions each calibrate the question differently. Provider-defendant cases can typically be brought in the provider's place of establishment; user-defendant cases in the user's domicile or in the place where the harmful conduct occurred or had its effects. The result is the multi-forum availability analyzed in §3.5.

Applicable law is governed in tort by Rome II (Regulation 864/2007) for EU fora, by analogous statutory or common-law rules elsewhere; in copyright by the *lex loci protectionis* principle of the Berne Convention; in contract by Rome I (Regulation 593/2008) and analogues; in data protection by the territorial-scope provisions of the relevant statutes. The plurality of applicable laws in a single AI-extensible product incident is the rule rather than the exception in cross-border cases.

Recognition and enforcement of judgments is governed for civil matters by the Hague Convention of 2 July 2019 on the Recognition and Enforcement of Foreign Judgments (entered into force 1 September 2023 for participating states), by bilateral and multilateral arrangements of greater coverage but more limited scope, and by the residual common-law and civil-law principles each forum applies. For criminal judgments, recognition is far more limited; the principal cooperative mechanisms are mutual legal assistance treaties (MLATs) and the transfer-of-sentenced-persons treaties.

The AI Act's extraterritoriality provisions (Art. 2) deserve specific note: the Act applies to providers placing AI systems on the EU market regardless of where the provider is established, to deployers located in the EU, and to providers and deployers located outside the EU where the output of the AI system is used in the EU. The reach echoes the GDPR's Art. 3(2) and adopts the same effective-jurisdiction strategy.

The practical consequence, traced in §3.6, is the "lowest-hanging-fruit" phenomenon: of the multiple substantively plausible cases, only those in which jurisdictional reach, evidentiary feasibility, and enforcement capacity converge actually proceed. The phenomenon raises rule-of-law concerns—defendants are sorted by reachability rather than culpability, victims recover unevenly across borders—that the substantive doctrines do not address and that the policy chapter takes up.

5.8 Other Doctrinal Implications

Several doctrinal fields that the article does not treat in depth are implicated by AI-extensible products and deserve brief flagging. *Labor law* engages questions of employer use of AI-extensible tools for productivity monitoring, performance assessment, and decision support; the EU AI Act classifies several such uses as high-risk and imposes corresponding obligations. *Consumer protection law* engages misleading representations about AI capabilities and unfair commercial practices in the deployment of AI in consumer-facing services. *Competition law* engages the position of dominant providers of foundation models on which AI-extensible products are built; the recent EU and US enforcement actions against vertical integration in the AI value chain are a developing front. *Securities and investment-advice regulation* engages AI-generated investment recommendations, where existing fiduciary and suitability frameworks must accommodate algorithmic origination. Each of these fields warrants its own sustained treatment, which exceeds the scope of the present article; each is identified here as an open question for subsequent scholarship.

6. Comparative Analysis Across Six Jurisdictions

6.1 Introductory Framing

The comparative analysis in this chapter examines six jurisdictions: Germany, Japan, China, the United States, the United Kingdom, and the European Union as a supranational regulatory regime. The selection is designed for diversity along the dimensions most relevant to the article's argument. Germany and Japan represent the continental civil-law tradition with strong doctrinal-systematic commitments, but operate in different cultural and institutional contexts. China combines a continental civil-law penal code with active judicial interpretation and policy steering of the kind familiar from socialist legal tradition. The United States represents the common-law tradition in its constitutionally-protected, intermediary-friendly variant. The United Kingdom represents post-Brexit common-law as it has begun to chart a distinct path between the US and EU positions. The EU is treated as a distinct unit of analysis because its supranational regulatory output is the most substantial in the AI-law field and because the interaction of EU regulation with member-state doctrine is itself a comparative-law phenomenon worth surfacing.

The analysis proceeds along four core doctrinal axes: criminal law, tort law, administrative regulation, and intellectual property. The remaining three doctrinal fields treated in Chapter 5 (contract, data protection, private international law) are referenced where they bear on the central axes but not given parallel jurisdictional treatment. Each jurisdictional section is organized around an *anchor case or statute*—a single judicial decision or legislative instrument that captures the jurisdiction's distinctive position—supplemented by the relevant secondary materials. The anchoring strategy avoids the encyclopedic trap of trying to summarize each jurisdiction's entire law on each doctrinal axis, while preserving sufficient specificity to support the synthesis at §6.8.

The comparative-law literature this chapter draws on is wider than the citations make explicit. Zweigert and Kötz's *Introduction to Comparative Law* and Patrick Glenn's *Legal Traditions of the World* supply the methodological framework; the field-specific comparative literatures (Markesinis and Unberath for tort, Weatherford and Pradi for criminal, Goldstein and Hugenholtz for copyright) supply the doctrinal substrate. The analysis aims at functional equivalence across systems while being attentive to the distinctive doctrinal architectures within which each system operates.

6.2 Germany

6.2.1 Governance Paradigm: Doctrinal Precision and Regulatory Embrace

The German governance posture toward AI-extensible products combines two features that elsewhere are often in tension. The first is the highly precise doctrinal analysis associated with the Roxin school of criminal law and the Larenz-Kötz tradition of civil law: the analysis of any AI-related conduct begins with the constitutive elements of the relevant offense or claim and proceeds through the established doctrinal apparatus of objective imputation, fault, and the like. The second is an active embrace of EU-level regulation: where the AI Act, the GDPR, the Product Liability Directive, and related instruments allocate competences to EU institutions, German implementation tends to be earnest and detailed rather than reluctant. The synthesis position has been particularly influential in the emerg-

ing German AI-law literature (Hilgendorf, Beck, Spindler), which combines doctrinal-criminal-law precision with engagement with the EU regulatory agenda.

6.2.2 Criminal Law: BGH *Mittäterschaft* Jurisprudence

The German criminal-law treatment of AI-extensible product harm is anchored in the doctrines of perpetration and complicity codified in §§25–27 StGB and elaborated in the Roxin tradition. §25(1) provides direct perpetration; §25(2) co-perpetration (*Mittäterschaft*); §§26 and 27 instigation and aiding. The BGH’s jurisprudence on *Mittäterschaft* requires *gemeinsamer Tatentschluss* (common plan) and substantial contribution to the joint execution (BGHSt 37, 289; BGHSt 49, 39; the more recent BGH 5 StR 65/19 and 2 StR 346/22 lines refining the contribution requirement). The common-plan element has been the principal point of doctrinal contention in multi-actor configurations.

Applied to the article’s scenarios: Scenario α ’s intentional sole user is reached through indirect perpetration using the AI as instrumentality, with no *Mittäterschaft* question. Scenario β ’s negligent user falls outside criminal fraud liability (no *fahrlässiger Betrug*); the only available criminal hooks are sectoral offenses keyed to specific publication of harmful content (some Computer-Strafrecht provisions, parts of §202c StGB on hacker tools), the application of which to general prompt-engineering material is contested. Scenario γ ’s multi-user configuration confronts the *Mittäterschaft* requirement: the BGH’s exacting view of common plan exonerates eA and eD from joint perpetration despite their parallel intent, leaving them as *Nebentäter* (concurrent perpetrators) each liable for her own contribution. The treatment of eB’s negligence again falls outside German fraud doctrine. eC is exonerated.

Hilgendorf’s *Robotik und Recht* treatments and his subsequent work on AI in the German Strafrecht series have argued for caution in extending the doctrinal apparatus to capture AI-mediated contributions; the proposed expansions of objective imputation to cover algorithmic causation that some commentators have suggested have not been adopted by the BGH, which has continued to apply established doctrine on its conventional terms. The German position on AI-related criminal liability is, accordingly, doctrinally parsimonious: the established categories apply where they apply, and conduct outside them is criminally non-labile, with the policy gap left to civil and administrative law.

6.2.3 Tort: BGB §823 and Product Liability

The German tort treatment proceeds principally through §823(1) BGB (injury to enumerated absolute rights), §823(2) (violation of a protective statute), and §826 (intentional conduct contrary to public morals). The protected interests include life, body, health, freedom, and “other rights” that German jurisprudence has elaborated to include the general right of personality (*allgemeines Persönlichkeitsrecht*) and the right to established and operating business (*Recht am eingerichteten und ausgeübten Gewerbebetrieb*). The duty-of-care standard is articulated through the doctrine of *Verkehrspflichten*: those who create or maintain sources of danger must take reasonable measures to prevent foreseeable harm to third parties.

The most important recent development for AI-extensible product liability is the 2024 revision of the Product Liability Directive (Directive (EU) 2024/2853, repealing Directive 85/374/EEC). The revised PLD extends product liability to standalone software and to AI-enabled products, addresses cybersecurity vulnerabilities as defects, and recasts the burden of proof in ways that ease the plaintiff’s

path. Implementation in German law (*Produkthaftungsgesetz*, currently being amended in line with the new Directive) makes Company B a meaningful target for civil recovery in a way that pre-2024 German law largely did not.

The proposed AI Liability Directive, which would have layered fault-based mechanisms with disclosure obligations and rebuttable causal presumptions onto the strict-liability framework of the PLD, has had a more difficult legislative path. At the time of writing the proposal has been withdrawn from the Commission's work programme; the question whether comparable mechanisms will be reintroduced in revised form, or whether national implementations of the revised PLD will fill the gap, remains open. The BGH has not yet had occasion to articulate AI-specific tort doctrines in the strict sense, but the cases working through the lower German courts on AI-generated defamation, AI-image misuse, and AI-mediated fraud are accumulating; the GEMA v. OpenAI litigation pending in Germany at the time of writing will likely produce the first authoritative German judgment on training-data uses that reach into the tort register.

The German approach to joint tortfeasance under §830 BGB has been discussed at §5.2; the treatment of Scenario γ under German tort doctrine reaches the intentional contributors and, on the prevailing view, leaves the negligent contributor liable under §823 if the duty-of-care analysis yields breach.

6.2.4 Administrative: AI Act Implementation

The German implementation of the EU AI Act, anticipated through a *Durchführungsgesetz* expected to be in force during 2025–2026, allocates competences among existing federal regulators rather than creating a single new AI authority. The Federal Office for Information Security (*BSI*) takes responsibility for AI system security and conformity assessment in security-related applications; sector-specific regulators (BaFin for finance, the Federal Network Agency for telecommunications, etc.) retain competence for AI deployments in their respective sectors. The *Bundesnetzagentur* is positioned to take broader coordination functions for general-purpose AI under the Act's distinctive framework.

The implementing legislation must navigate several substantive questions left open by the Act itself: the precise content of high-risk-system risk-management obligations as concretized for particular sectors; the interaction between the Act's transparency obligations and the German constitutional protection of trade secrets; the procedure for conformity assessment and market surveillance; the calibration of administrative penalties within the Act's permitted ranges. The earlier German engagement with the GDPR implementation (*BDSG-neu*) provides a template for the kind of thoroughgoing implementation expected.

The implementation's significance for AI-extensible products is twofold. First, providers offering products in the German market face the AI Act's regulatory architecture in its German concretization, which is likely to be more demanding than the Act's text alone suggests. Second, the German enforcement experience, given the BSI's technical capacity and the regulator-civil-society engagement historically prevalent in German data-protection enforcement, will be a leading indicator of how the Act will be operationalized across the EU. The article returns to the cross-cutting AI-Act analysis in the EU subsection at §6.7.

1927 6.2.5 IP: §69 Series and AI Training

1928 German copyright law's distinctive features for AI-extensible product analysis are concentrated in
 1929 two provisions of the *Urheberrechtsgesetz* (UrhG). The §69 series (§§69a–69g) implements the EU
 1930 Software Directive's special rules for computer programs, providing the basis for protection of the
 1931 provider's underlying software and configuration. §44b, implementing the EU DSM Directive's
 1932 text-and-data-mining exception, permits text and data mining for any purpose subject to a machine-
 1933 readable opt-out by the rightholder; §60d extends the permission for non-commercial research. The
 1934 opt-out provision is the central mechanism by which German copyright holders can prevent the use
 1935 of their works for AI training, and its operation has been at the centre of recent litigation.

1936 The Hamburg Regional Court's judgment in *Kneschke v. LAION* (LG Hamburg, Urt. v. 27.9.2024
 1937 – 310 O 227/23) addressed the applicability of §60d to a research-purposed image dataset; the court
 1938 held the use covered by the research exception, in a decision that has been widely read as supportive
 1939 of training-data use within research-coded uses but not yet authoritatively extended to commercial
 1940 training. The pending GEMA litigation against OpenAI (filed November 2024) raises the broader
 1941 question whether commercial generative-AI training on copyrighted works without opt-out is per-
 1942 mitted under §44b and the DSM Directive analysis. The CJEU's eventual jurisprudence on these
 1943 provisions will substantially shape the German position.

1944 The authorship analysis for AI-extensible product outputs follows the EU-level position discussed
 1945 at §5.3: human creativity is required for copyright subsistence, with mixed human-AI works analysed
 1946 contribution-by-contribution. The German literature on this question (Lauber-Rönsberg, Hetmank,
 1947 Apel) has elaborated the detailed application but has not departed significantly from the EU position.

1948 6.3 Japan

1949 6.3.1 Governance Paradigm: Doctrinal Restraint and Soft-Law Preference

1950 The Japanese governance posture combines two features that, while separately present in other juris-
 1951 dictions, distinguish the Japanese position when present together. The first is the Hirano-Matsumiya
 1952 tradition of 謙抑主義 (restraint principle) in criminal law, which holds that criminal liability should
 1953 be expanded with extreme caution and that absent clear statutory mandate the doctrine should err on
 1954 the side of non-criminalization. The tradition is visible in the BGH-comparable conservatism of the
 1955 Supreme Court of Japan on questions of complicity and indirect perpetration, and in the influential
 1956 commentary of 松宮孝明, 佐伯仁志, and 山口厚. The second is the preference for soft-law reg-
 1957 ulation: the principal Japanese regulatory output on AI consists of ministry-issued guidelines (ガイ
 1958 ドライン) rather than binding statutes, supplemented by industry self-regulation and the expansion
 1959 of existing sectoral frameworks. The result is an AI legal landscape that is doctrinally familiar but
 1960 normatively elastic, with extensive room for case-by-case adaptation but limited ex-ante clarity.

1961 6.3.2 Criminal Law: The Winny Doctrine

1962 The anchor judgment for Japanese criminal-law analysis of provider-mediated harm is the Supreme
 1963 Court's decision in the *Winny* case (最決平成 23 年 12 月 19 日刑集 65 卷 9 号 1380 頁). The
 1964 defendant had developed and distributed the Winny peer-to-peer file-sharing software, which was

widely used for copyright infringement in addition to lawful uses. The first-instance court convicted him of 著作権法違反幫助 (complicity in copyright infringement); the Osaka High Court reversed; the Supreme Court affirmed the High Court's reversal and articulated the test now known as the *Winny* doctrine.

The doctrine has two prongs. First, a person who provides software that has both lawful and unlawful uses (an 価値中立的な技術, value-neutral technology) is liable for complicity in users' infringement only if she possessed the specific subjective awareness that the software would be used predominantly for infringing purposes. Second, the standard for "predominantly" is exacting: the prosecution must establish that the unlawful use was, in the provider's contemporaneous understanding, the principal purpose for which she was making the software available. The Supreme Court emphasized the negative consequences for technological innovation of a more permissive standard.

Applied to AI-extensible products: Company B is offering a value-neutral technology in the relevant sense; the lawful uses are substantial; and the prosecution would have great difficulty establishing that Company B's subjective understanding was that the principal purpose of its product was unlawful. The *Winny* doctrine accordingly exonerates the provider in the range of cases the article analyzes. Scenario α 's intentional user is liable through indirect perpetration or as direct perpetrator of 詐欺罪 (Penal Code Art. 246); Scenario γ 's intentional users are liable individually but not as joint perpetrators in the absence of 意思の連絡; the negligent and innocent users are non-liable; the provider is non-liable.

The Japanese commentary (松宮, 刑法総論講義; 佐伯仁志, 刑法総論の考え方・楽しみ方; 山口厚, 刑法総論) has generally endorsed the *Winny* approach as consistent with the 謙抑主義 principle, while acknowledging the cost in provider non-attribution for foreseeable but not specifically-intended harms. The doctrinal stability the *Winny* approach provides may be tested by future AI-related cases in which the harms are more severe and the provider's foreseeability of the use category is harder to deny.

6.3.3 Tort: Civil Code Art. 709 and Network Infringement

The Japanese tort treatment of AI-extensible product harm is anchored in Civil Code Art. 709 (general tort liability for intentional or negligent infringement) and Art. 719 (joint torts and uncertain attribution among multiple actors), supplemented by the プロバイダ責任制限法 (Provider Liability Limitation Act, 2001, materially amended in 2021 and renamed 情報流通プラットフォーム対処法 in 2024).

The general Art. 709 analysis follows the framework outlined in §5.2: intentional contributors to harm are liable; negligent contributors are liable if their conduct breached a duty of care owed to the victim; innocent contributors are non-liable. The duty-of-care analysis under Japanese tort law is contested in the AI context. The judgments closest to the AI-extensible product configuration come from intermediary-liability jurisprudence developed in pre-AI network-infringement cases (the ニフティサーブ line of decisions, the *TVブレイク* judgments, and the recent generative-AI defamation cases beginning to be brought in the Tokyo District Court). These judgments calibrate provider duty by reference to the provider's practical capacity to prevent harm and the social value of the service offered.

2005 The 2024 情報流通プラットフォーム対処法 extended and clarified the provider notice-
 2006 and-takedown framework, introduced faster takedown timelines for specified categories of content,
 2007 and clarified provider obligations with respect to user identification for tort-victim plaintiffs. The
 2008 Act applies to “information- distribution platforms” as defined; whether AI-extensible products fall
 2009 within the definition is contested and likely jurisprudentially unsettled until the Supreme Court rules.

2010 The joint-tortfeasance analysis under Art. 719 has been the subject of extensive Japanese schol-
 2011 arship; the prevailing view (潮見, 窪田) treats the first sentence’s 共同 requirement as an objective
 2012 relatedness criterion that can be satisfied by interaction through a shared intermediary process. On
 2013 this view, the multi-user contributors of Scenario γ could be held jointly and severally liable to V
 2014 even in the absence of subjective coordination. The point has not been authoritatively decided in an
 2015 AI context.

2016 **6.3.4 Administrative: AI 事業者ガイドライン**

2017 The principal Japanese administrative-law instrument for AI is the *AI 事業者ガイドライン* (AI
 2018 Business Operators’ Guidelines) issued jointly by the Ministry of Economy, Trade, and Industry (経
 2019 済産業省) and the Ministry of Internal Affairs and Communications (総務省) in April 2024 and
 2020 updated periodically. The Guidelines consolidate three earlier sets of guidance (the AI Development
 2021 Guidelines of 2017, the AI Utilization Guidelines of 2019, and the Governance Guidelines for Imple-
 2022 mentation of AI Principles of 2022) into a unified framework addressing “business operators” broadly
 2023 conceived: developers, providers, and users.

2024 The Guidelines are explicitly soft-law: compliance is voluntary, no administrative penalties at-
 2025 tach to non-compliance, and the instrument’s normative force depends on industry uptake and on
 2026 the indirect channels through which guidelines influence litigation (reasonable-care standards, reg-
 2027 ulatory expectations in adjacent domains). The Japanese preference for soft-law regulation reflects
 2028 several considerations: the institutional preference for industry-government consensus-building (審
 2029 議会); the constitutional allocation of legislative authority that makes binding regulation a multi-year
 2030 process; and a substantive judgment that AI is too rapidly changing for binding regulation to keep
 2031 pace.

2032 The cost of the soft-law approach is the absence of clear ex-ante obligations and the reliance on
 2033 civil and reputational mechanisms for compliance pressure. The benefit is flexibility and the avoid-
 2034 ance of regulatory ossification. The article does not take a position on whether the Japanese trade-off
 2035 is on balance correct; it observes that the trade-off positions Japanese AI-extensible product gover-
 2036 nance at a substantially different point on the regulatory spectrum than the EU and China positions,
 2037 and that Japanese providers and users operate under correspondingly different ex-ante constraints.
 2038 The shift to binding regulation, contemplated in the discussions surrounding the proposed AI promo-
 2039 tion legislation under consideration in the Diet at the time of writing, may move the position closer
 2040 to the EU end of the spectrum.

2041 **6.3.5 IP: Copyright Act Art. 30-4**

2042 The most distinctive feature of Japanese copyright law for AI-extensible product analysis is the Copy-
 2043 right Act 第 30 条の 4 (Art. 30-4), introduced in the 2018 amendment. The provision permits the use

of copyrighted works for purposes that do not involve the enjoyment of the work's expressive value, and expressly enumerates "information analysis" (情報解析) as a covered purpose. The provision permits use "to the extent considered necessary," subject to a proviso that the use must not unreasonably prejudice the interests of the rightholder in light of the nature, purpose, or mode of the work's use.

The 2018 amendment was deliberately drafted to be permissive of AI training, and the Japanese position is consequently the most permissive among the major jurisdictions surveyed. AI training on copyrighted works that does not involve enjoyment of expression—a category typically read to include the training of generative models—is permitted without licence and without opt-out. The Cultural Affairs Agency's 2024 *AI と著作権に関する考え方について* provides authoritative guidance on the application of Art. 30-4 to generative AI; the guidance distinguishes between training (covered by Art. 30-4) and generation (governed by ordinary copyright analysis, with infringement turning on substantial similarity and the dependence relationship).

The permissive Japanese position has shaped the AI ecosystem in distinctive ways. Several major Japanese AI developers have located training operations in Japan to take advantage of Art. 30-4; foreign providers operating in the Japanese market enjoy the same advantage. The provision is at the same time a source of friction with foreign copyright holders, particularly content industries based in jurisdictions with more restrictive training regimes. The CJEU's forthcoming judgments on the EU DSM Directive's TDM exception, and the US courts' resolution of pending generative-AI copyright litigation, will determine whether the Japanese position remains an outlier or becomes part of an emerging international pattern.

The authorship analysis for AI-extensible product outputs follows the framework of §5.3: human creative contribution is required for copyright subsistence, and the Cultural Affairs Agency's 2024 guidance applies a contribution-by-contribution analysis to mixed human-AI works. The Japanese position on output copyright is, accordingly, less distinctive than its position on training-data uses.

6.4 China

6.4.1 Governance Paradigm: Codified Doctrine, Active Judicial Interpretation, Policy Steering

The Chinese governance posture combines a continental-style codified substantive law (Criminal Law of 1997, with multiple amendments; Civil Code of 2020 consolidating prior dispersed civil legislation; Copyright Law of 1990, last amended 2020) with the distinctive mechanism of judicial interpretation (司法解释) by which the Supreme People's Court and the Supreme People's Procuratorate—the "Two Highs" (两高)—issue authoritative interpretations that bind lower courts. Judicial interpretations carry quasi-legislative weight in the operative system and have been the principal vehicle for elaborating the application of code provisions to new technologies. The third feature is policy-driven enforcement prioritization: CCP and State Council policy documents identify priority areas, and prosecutorial and regulatory enforcement concentrates on those areas in pulses that are visible across the enforcement record.

AI regulation has been a notably proactive enforcement priority. The Cyberspace Administration

of China (CAC, 国家互联网信息办公室) issued generative-AI service regulations in July 2023 that predate the EU AI Act by nearly a year and impose obligations comparable in some respects to the AI Act's high-risk provisions. The regulatory output continues with algorithmic recommendation regulations (2022), deepfake regulations (2023), and the contemplated comprehensive AI Law (drafts circulated 2024–2026). The combined effect is a body of binding AI-specific administrative law that is, in coverage and detail, comparable to the EU's.

6.4.2 Criminal Law: Aiding-Information-Crime Offense

The most distinctive element of Chinese criminal-law treatment of provider-mediated harm is the offense of 帮助信息网络犯罪活动罪 (Aiding Information Network Criminal Activity, Criminal Law Art. 287-2, often abbreviated as 帮信罪), introduced in the 2015 Criminal Law Amendment (Ninth). The offense criminalizes the knowing provision, to a person committing an offense using an information network, of internet access, server hosting, network storage, communications transmission, advertising promotion, payment settlement, or other technical assistance. The knowledge element (明知) was clarified by the Two Highs' Judicial Interpretation of 25 October 2019 on Several Issues Concerning the Application of Law in Handling Criminal Cases of Aiding Information Network Criminal Activity, which permitted the inference of knowledge from objective factors including transaction patterns, repeat dealings, abnormal compensation arrangements, and continuation after warning.

The offense has been used extensively in cyber-fraud enforcement since 2019, with conviction figures in the hundreds of thousands. Its application to AI-extensible product providers is, at the time of writing, contested and not yet anchored in a leading judgment. The doctrinal availability is clear: a provider who knowingly provides AI services that are used for fraud satisfies the constitutive elements, with the knowledge element subject to the 2019 Interpretation's inference rules. The prosecutorial discretion that conditions actual application is the principal constraint.

The comparative significance of 帮信罪 is that it offers a low-threshold attribution tool for provider liability that neither the German doctrinal apparatus nor the Japanese *Winny* doctrine offers. Where the *Winny* doctrine requires that unlawful use be the principal purpose of the provider's distribution, 帮信罪 requires only knowledge that the service is being used for offense and continued provision. Where German doctrine requires *Mittäterschaft* or *Beihilfe* elements that AI providers do not satisfy, 帮信罪 treats the provider as a directly punishable principal of a distinct offense. The result is that the same provider, on the same conduct, faces materially different criminal exposure across the three jurisdictions—a point developed at §7.4.

The leading academic commentary (张明楷, 陈兴良, 周光权) has debated the proper scope of 帮信罪 and the application of the 2019 Interpretation; the dominant view supports a substantial scope but counsels prosecutorial restraint to avoid the over-criminalization the offense's elements would otherwise permit. Whether the restraint will be observed in AI-related applications is uncertain.

6.4.3 Tort: Civil Code Art. 1194–1197

The principal Chinese tort provisions for AI-extensible product analysis are concentrated in Civil Code Articles 1194–1197, which elaborate the liability of network service providers and users. Art. 1194

imposes general liability on a network user or provider who infringes the civil rights and interests of others through the use of a network. Art. 1195 establishes the notice-and-takedown framework: rightholders may notify the network service provider, which must take necessary measures (deletion, blocking, link removal); the rightholder bears liability for false notices. Art. 1196 imposes joint and several liability where the provider knew or should have known of the infringement and failed to take necessary measures. Art. 1197 addresses the special case where the rightholder cannot identify the infringing user and seeks the provider's assistance in identification.

The framework was drafted before the widespread deployment of generative AI and presupposes a paradigm in which the provider's role is the hosting or transmission of user-generated content rather than the active generation of content through user interaction with an embedded model. Application to AI-extensible products is in flux. Several lower-court judgments have applied the framework to AI-image and AI-text infringement cases by analogy, treating the provider as a network service provider for purposes of the notice-and-takedown obligations and the constructive-knowledge liability of Art. 1196. Whether the framework's notice-and-takedown structure can effectively address harms that arise from generative outputs (rather than persistent user-uploaded content) is the principal open question.

The interaction with the criminal-law framework of §287-2 deserves note. The same provider conduct may attract civil liability under Art. 1196 (constructive knowledge plus failure to take measures) and criminal liability under 帮信罪 (subjective knowledge plus continued provision). The Chinese civil-criminal doctrinal articulation does not require either to follow from the other, but the evidentiary thresholds interact in practice: a successful 帮信罪 prosecution typically supports civil liability; the converse is not true.

6.4.4 Administrative: Generative AI Interim Measures

The principal administrative-law instrument for generative AI in China is the 生成式人工智能服务管理暂行办法 (Generative AI Service Management Interim Measures), jointly issued by seven agencies under CAC's leadership in July 2023 and effective from August 2023. The Measures impose obligations on providers of generative AI services to the Chinese public including: registration with the CAC under the security-assessment regime (算法备案); compliance with content-moderation obligations keyed to the same prohibited-content categories that govern other internet services in China; data-quality and privacy obligations in respect of training data; transparency obligations including labelling of AI-generated content; and liability for unlawful outputs traceable to inadequate training-data curation or inadequate ex-post controls.

The Measures supplement, and operate against the background of, the 互联网信息服务算法推荐管理规定 (Internet Information Service Algorithmic Recommendation Provisions, 2022) and the 互联网信息服务深度合成管理规定 (Deep Synthesis Provisions, 2022), each of which addresses a particular slice of algorithm-mediated information services. The contemplated comprehensive AI Law, drafts of which have been circulating since 2023, would consolidate and extend these instruments; whether the final version retains the current Measures' approach or adopts elements of the EU AI Act's risk-tier architecture is an open question.

The Chinese administrative regime is the most developed body of binding AI-specific law among

the jurisdictions surveyed. Its distinctive features include the integration of content-moderation obligations with technical-safety obligations (a combination the EU AI Act largely separates between the AI Act and the Digital Services Act), the strong role of the CAC as a single coordinating authority, and the linkage to the broader cybersecurity and data-security legislative architecture (the Cybersecurity Law of 2017, the Data Security Law of 2021, the PIPL of 2021). Providers operating in China face a more comprehensive ex-ante regime than in any other jurisdiction surveyed, with correspondingly substantial compliance burdens but also relative clarity on the regulatory expectations.

6.4.5 IP: Beijing Internet Court AI Image Case

The Beijing Internet Court's judgment in *Li v. Liu* of 27 November 2023 ((2023) 京 0491 民初 11279 号) is the leading Chinese authority on copyright in AI-assisted output. The plaintiff had used the Stable Diffusion model to generate an image, supplying the prompts, parameter selections, and post-generation modifications. The defendant subsequently used the image online without permission. The court held that the plaintiff's contribution—in selecting and refining prompts, choosing parameters, and post-processing—constituted creative intellectual labor sufficient for copyright subsistence; the image was protected as a work of fine art with the plaintiff as author.

The judgment is the most plaintiff-favorable AI-authorship judgment among the jurisdictions surveyed. It diverges from the US position under *Thaler v. Perlmutter* and Copyright Office guidance, which limits copyright in AI-generated material to elements of human authorship narrowly defined. It diverges from the German and EU positions, which require “author's own intellectual creation” in a sense that, as currently elaborated, would not be satisfied by prompt-based generation alone. It is closer to but more permissive than the UK's §9(3) Copyright, Designs and Patents Act provision for computer-generated works, which has not been authoritatively applied to modern AI.

The judgment's reach is contested within China. Subsequent decisions of other intermediate-level courts have followed the Beijing Internet Court's approach in some cases and resisted it in others. The Supreme People's Court has not yet authoritatively resolved the question. The article does not predict whether the *Li v. Liu* approach will become the consolidated Chinese position or whether the contested status will continue; it observes that the divergence from the US/EU positions is itself a comparative phenomenon that the synthesis at §6.8 engages.

China's broader copyright treatment of AI training data is in formation. There is no Chinese analog to Japan's permissive Art. 30-4 or to the EU DSM Directive's TDM exception with opt-out; the principal Chinese training-data uses operate on the basis of fair-use-style exceptions in Copyright Law Art. 24 and on implicit toleration where the training is conducted by domestic champions. The recent litigation against generative-AI providers in China (the Guangzhou and Shanghai Internet Courts have heard several such cases in 2024–2025) will begin to articulate the operative position.

6.5 United States

6.5.1 Governance Paradigm: Constitutional Protection of Intermediaries

The US governance posture toward AI-extensible products is shaped by three structural features that distinguish it from the other jurisdictions surveyed. The first is the First Amendment's protection

of speech and press, which has been read to constrain government regulation of online content and through the line of cases beginning with *Reno v. ACLU* (1997), of internet intermediaries. The second is Section 230 of the Communications Decency Act of 1996 (47 USC §230), which provides broad immunity to providers and users of interactive computer services from liability for content provided by third parties. The third is the generally permissive regulatory posture toward AI development, reflected in the National Institute of Standards and Technology (NIST) AI Risk Management Framework’s voluntary character and the absence of comprehensive federal AI legislation.

The combination makes the United States the most provider-friendly jurisdiction surveyed. Providers operating in the US face less exposure to civil liability for user-mediated harm than in any other jurisdiction; less ex-ante regulatory burden than in the EU or China; and the protection of constitutional speech doctrines that limit congressional and state response to AI-mediated harms. The corresponding gap on the victim side—third parties harmed through provider-mediated activity have substantially fewer remedies than in EU or Chinese fora—has been at the centre of recent legislative and judicial debate.

6.5.2 Criminal Law: MPC §2.02 and *Rosemond*

US criminal-law analysis of AI-mediated harm proceeds in the federal system through the offense-specific provisions of Title 18 (most prominently the wire-fraud statute, 18 USC §1343) and the aiding-and-abetting statute (§2); in state systems through penal codes derived in varying degrees from the Model Penal Code (MPC). The MPC §2.02 four-tier mental-state hierarchy (purposely, knowingly, recklessly, negligently) and §2.06 on complicity supply the canonical doctrinal structure for the roughly two dozen states that have substantially adopted the MPC, and influence the analysis even in non-MPC states.

The leading recent Supreme Court decision on accomplice liability is *Rosemond v. United States*, 572 U.S. 65 (2014), which held that conviction for aiding and abetting under 18 USC §2 requires that the defendant have advance knowledge of all of the elements of the underlying offense. Applied to the article’s scenarios: Scenario α ’s intentional sole user is liable as principal under the wire-fraud statute (the deceptive transmission satisfies the elements; the AI is instrumentality). Scenario β ’s negligent user is non-labile for the intentional fraud offense and unlikely to fall under any sectoral or strict-liability hook. Scenario γ ’s two intentional users (eA, eD) are individually liable as principals; the question whether they are co-conspirators or aiders-and-abettors of one another turns on whether the forum-based pseudonymous coordination satisfies the agreement element of conspiracy or the advance-knowledge element of *Rosemond*. The case for either is uncertain; conviction as sole intentional perpetrators is, however, secure on the facts. eB’s negligence and eC’s innocence place them outside criminal liability.

The wire-fraud statute’s substantive scope is broad and its extraterritorial reach is expansive: it has been applied to foreign-located conduct that uses US wire infrastructure or that targets US-located victims. Applied to Scenario ε , the United States has plausible jurisdiction over eD (US national, US conduct) and arguably over the provider Company B (Delaware incorporation), with the latter’s exposure substantially limited by the Section 230 analysis of the next subsection. The federal prosecution practice in cyber-fraud cases is increasingly active; generative-AI-enabled fraud is an

2240 explicit DOJ enforcement priority since 2024.

2241 **6.5.3 Tort: Section 230 and Gonzalez**

2242 US tort treatment of AI-extensible product harm is dominated by the operation of Section 230 of the
 2243 Communications Decency Act (47 USC §230). The statute provides that “[n]o provider or user of an
 2244 interactive computer service shall be treated as the publisher or speaker of any information provided
 2245 by another information content provider.” The provision has been read by US courts since *Zeran v.*
 2246 *America Online* (4th Cir. 1997) to confer broad immunity on intermediaries for content created by
 2247 third parties.

2248 The application of Section 230 to AI-generated content is the defining open question for US AI-
 2249 tort doctrine. *Gonzalez v. Google LLC*, 598 U.S. 617 (2023), presented the Supreme Court with the
 2250 question whether algorithmic recommendations of third-party content fall within Section 230’s pro-
 2251 tection. The Court declined to reach the merits, remanding on companion-case grounds and explicitly
 2252 leaving the algorithmic question for another day. The lower-court litigation since *Gonzalez* has begun
 2253 to address the cognate but distinct question whether content generated by a provider’s AI system—
 2254 not merely recommended by an algorithm but produced through it— is “information provided by
 2255 another information content provider” or content for which the provider is itself the “information
 2256 content provider” under Section 230(f)(3) and thus outside the immunity.

2257 Volokh’s treatment in *Large Libel Models* (2023) argues that generated content for which the
 2258 user supplied only a brief prompt is most plausibly the provider’s content for Section 230 purposes,
 2259 exposing the provider to defamation liability under ordinary common-law principles. The compet-
 2260 ing position—that generated content remains user-attributable when the user’s prompt initiated its
 2261 production—has been advanced by provider-side commentators and in litigation positions but has
 2262 not been authoritatively adopted. The article does not predict the resolution; it observes that the
 2263 resolution will substantially shape the US tort exposure of AI-extensible product providers, with
 2264 corresponding effects on provider design choices and on the comparative position of US providers
 2265 vis-à-vis providers in regulatory regimes that do not turn on the Section 230 question.

2266 Beyond the Section 230 question, US tort doctrine on AI-mediated harm proceeds through ordi-
 2267 nary negligence, products liability, and intentional-tort frameworks. The products-liability analysis
 2268 is in flux: the Restatement (Third) of Torts: Products Liability, drafted in the 1990s, treats software
 2269 and AI on terms that the contemporary practice has begun to depart from. The federal courts have
 2270 not yet authoritatively addressed whether AI-extensible products are “products” for products-liability
 2271 purposes, but the trend in state-court adjudication has been toward inclusion in some respects and not
 2272 in others.

2273 **6.5.4 Administrative: NIST AI RMF and Sectoral Regulation**

2274 US administrative regulation of AI proceeds through three principal channels: voluntary federal
 2275 frameworks, sector-specific federal regulation, and state legislation. The NIST AI Risk Management
 2276 Framework (RMF), released in January 2023 with a Generative AI Profile in July 2024, provides a
 2277 voluntary structure for organizational risk management. The RMF identifies four functions (Govern,
 2278 Map, Measure, Manage) and elaborates each in detailed guidance; it imposes no binding obligations

and conditions no penalties on non-compliance. Federal agency compliance with the RMF is mandated by Office of Management and Budget guidance for federal AI deployments, but the framework does not extend to private-sector deployments except through indirect channels (federal procurement, liability standards, regulatory expectations).

The sectoral federal channels include the FTC’s enforcement against deceptive and unfair practices in AI deployment under 15 USC §45; the SEC’s enforcement against AI-related disclosure violations and fraud; the EEOC’s guidance on AI in employment decisions; the FDA’s regulation of AI in medical devices; the NHTSA’s regulation of AI in autonomous vehicles. These channels address particular slices of AI deployment without providing a horizontal framework.

The state channels have been the most active legislative front in 2024–2026. California’s AI Transparency Act, the Colorado Artificial Intelligence Act of 2024 (the first comprehensive state AI statute, effective February 2026), the Texas Responsible AI Governance Act, and the New York City automated employment decision-making rules collectively constitute a developing patchwork of state-level regulation. The Colorado Act in particular adopts a risk-tier approach broadly comparable to the EU AI Act’s high-risk category, with corresponding compliance obligations for “developers” and “deployers” of high-risk AI systems.

The patchwork raises distinctive questions of preemption, extraterritorial reach, and forum-shopping. The Trump administration’s January 2025 executive order revoking the Biden-era AI executive order signaled a shift toward state-led regulation in lieu of comprehensive federal action; whether the state patchwork will consolidate around a common framework, fragment further, or be displaced by federal preemption remains unsettled.

6.5.5 IP: Copyright Office Guidance and *Thaler*

US copyright treatment of AI-extensible product output is anchored in *Thaler v. Perlmutter*, 687 F. Supp. 3d 140 (DDC 2023), *aff’d*, 130 F.4th 1 (DC Cir. 2025), and in the Copyright Office’s three-part guidance on copyright registration involving AI (beginning with the March 2023 statement of policy and continuing with the Office’s reports on copyrightability of January 2025 and on training data of December 2025).

Thaler confirmed the Copyright Office’s longstanding position that copyright requires human authorship: a work autonomously generated by an AI system without sufficient human creative contribution is not copyrightable. The DC Circuit’s affirmance grounds the holding in the statutory text of the Copyright Act and in the historical understanding of authorship reflected in Supreme Court jurisprudence from *Burrow-Giles* (1884) forward.

The Copyright Office’s guidance has elaborated the application to mixed human-AI works through registration practice. The leading registration decisions (*Zarya of the Dawn*, 2023; the *Théâtre D’opéra Spatial* denial; the partial registration practice for AI-assisted works with substantial human-authored arrangement) establish that copyright subsists in the human-authored elements but not in elements “generated by AI” in the relevant sense. The Office has applied a substantial-human-contribution standard that is more demanding than the Beijing Internet Court’s standard but less demanding than a pure human-only threshold.

The bright-line rule against AI authorship has substantial practical implications for AI-extensible

product users: outputs of the kind the products generate are largely uncopyrightable in the United States, with the corresponding consequence that competitors can freely copy them. Users seeking copyright protection for AI-assisted output must structure the production to ensure substantial human authorship of the protected elements. Provider ToS that purport to assign output copyright to the user may convey nothing where no copyright subsists.

The pending generative-AI training-data litigation in US federal courts (the New York Times case against OpenAI; the Authors Guild class actions; the Andersen artists' suit against Stable Diffusion; the music-publisher cases) addresses the parallel question of whether unlicensed training on copyrighted works infringes. The defendants principally rely on fair use under 17 USC §107; the plaintiffs distinguish prior fair-use technology cases on the output-substitution dimension. The cases are at various procedural stages and will substantially shape the US position on training-data uses.

6.6 United Kingdom

6.6.1 Governance Paradigm: Common Law Plus Active Statutory Reform

The UK governance posture, in the post-Brexit period, has charted a position distinct from both the US and the EU. The platform-liability position established by the Online Safety Act 2023 is materially more aggressive than the US Section 230 baseline; the AI-specific regulatory position articulated in the AI Regulation White Paper of March 2023 and elaborated through subsequent agency-led activity is materially less prescriptive than the EU AI Act's. The combination—comprehensive duties of care for online platforms, with sectoral and principles-based regulation of AI specifically—is unusual and reflects the UK's deliberate post-Brexit positioning as a “pro-innovation” regulator while preserving substantial victim protection in the platform context.

The post-Brexit divergence from EU law has been incremental rather than dramatic. The UK GDPR retains substantial alignment with the EU GDPR; the UK PLD (Consumer Protection Act 1987 with anticipated amendments) is being updated in parallel with but not identical to the EU PLD revision; the UK AI Safety Institute (established 2023, since renamed AI Security Institute) operates in parallel with the EU AI Office. The Labour government elected in July 2024 has indicated continued commitment to the principles-based approach while signaling more active regulatory engagement on foundation-model deployment.

6.6.2 Criminal Law: *Jogee* and Common Law Complicity

The leading recent UK Supreme Court authority on accessory liability is *R v Jogee* [2016] UKSC 8, which reformulated the common-law doctrine of joint enterprise. Before *Jogee*, the “parasitic accessory liability” doctrine of *Chan Wing-Siu v R* [1985] AC 168 had imposed liability on a secondary party for the principal's offense whenever the secondary party foresaw the possibility that the principal might commit it. *Jogee* abolished this branch of the doctrine, holding that secondary liability requires that the secondary party intentionally assist or encourage the principal's commission of the offense, with knowledge of the essential facts. Foreseeability of the offense as a possibility is, after *Jogee*, evidence of intent but not itself sufficient.

Applied to AI-mediated cooperation: Scenario α 's intentional sole user is liable as principal for

fraud (the Fraud Act 2006). Scenario β 's negligent user is non-labile for fraud (no negligent variant) and unlikely to fall under any ancillary offense; the ONLINE SAFETY ACT's priority offenses for criminal content do not reach negligent prompt publication absent specific illegality of the content category. Scenario γ 's two intentional users are individually liable; the question whether they are co-accessories or jointly engaged requires intentional assistance or encouragement under *Jogee*, which on the facts of pseudonymous forum coordination is at most weakly satisfied.

The provider's position under UK criminal law is governed primarily by the underlying offense's elements and by the OSA's specific offenses (treated under the administrative subsection below). The common-law accessorial standard of *Jogee*, applied to the provider, would require intentional assistance with knowledge of essential facts—a standard the provider does not satisfy in the typical AI-extensible product configuration. The cumulative result is closer to the German/Japanese pattern of provider non-attribution than to the Chinese pattern of ready provider attribution under 帮信罪.

Smith & Hogan, *Criminal Law* (16th ed. 2024) treats the *Jogee* reformulation as having clarified rather than radically transformed accessorial doctrine; the practical result in fraud cases has been a narrowing of secondary liability that the article's analysis treats as broadly continuous with the post-2016 trajectory.

6.6.3 Tort: Common Law Negligence and OSA

UK tort treatment of AI-extensible product harm proceeds through common-law negligence (with the *Caparo* three-stage test for duty—foreseeability, proximity, and the fair, just and reasonable inquiry); through the law of nuisance and trespass for particular harms; and through the developing duty-of-care framework under the Online Safety Act 2023 for online services within the Act's scope. The 2024 *Hill v Smith* line and subsequent decisions have begun to address AI-specific duty-of-care questions, but the field is in formation.

The most consequential recent development is the OSA's imposition of specific statutory duties of care on online service providers in respect of illegal content, content harmful to children, and specified priority offenses. The duties include risk assessment, content removal, transparency reporting, and user-empowerment obligations; Ofcom is empowered to enforce through penalties of up to ten percent of global turnover. The interaction with common-law negligence is subtle: the OSA does not displace the common law but creates a parallel statutory duty, with the common-law analysis informed by the OSA's specifications of expected provider conduct.

Whether AI-extensible products fall within the OSA's "user-to-user service" or "search service" definitions is contested. Ofcom's guidance has indicated that some AI-mediated services fall within scope; the precise boundary is being worked out through enforcement practice. To the extent AI-extensible products are within scope, providers face statutory duties substantially more demanding than the US Section 230 baseline; to the extent they are outside, the common-law negligence framework applies in its less specified form.

The product-liability dimension is governed by the Consumer Protection Act 1987, which is being amended in light of the EU PLD revision although on a UK-specific timeline. The Act's coverage of software and AI is in flux; the post-Brexit divergence from the EU PLD may produce material differences in coverage by the late 2020s.

2397 **6.6.4 Administrative: OSA 2023 and AI White Paper**

2398 The UK administrative regulation of AI-extensible products operates through two principal channels.
 2399 The Online Safety Act 2023, in force from 2024, imposes the duties of care described above on online
 2400 services within scope, with Ofcom as enforcing authority. The AI Regulation White Paper of March
 2401 2023, “A pro-innovation approach to AI regulation,” articulated the UK government’s preference
 2402 for principles-based, sector-led regulation in lieu of a horizontal AI act of the EU type. The White
 2403 Paper identified five principles—safety, security, and robustness; transparency and explainability;
 2404 fairness; accountability and governance; contestability and redress—to be applied by existing sectoral
 2405 regulators (the FCA, MHRA, Ofcom, ICO, etc.) within their respective jurisdictions.

2406 The principles-based approach was implemented through the issuance of sectoral guidance, the
 2407 establishment of the AI Safety Institute (now AI Security Institute) within DSIT for evaluation of
 2408 frontier foundation models, and the negotiation of voluntary commitments from leading providers (the
 2409 Bletchley commitments of November 2023 and the Seoul commitments of May 2024). The Labour
 2410 government has indicated intent to introduce binding regulation of foundation models specifically,
 2411 while preserving the broader principles-based approach for downstream AI applications. The shape
 2412 of the contemplated AI Bill remains in development at the time of writing.

2413 The result is a regulatory landscape that is, in its current form, substantially less prescriptive
 2414 than the EU AI Act but materially more interventionist than the US federal baseline. Providers op-
 2415 erating in the UK face the OSA’s duties of care for in-scope services and the sectoral regulators’
 2416 principles-based expectations for AI deployments; the absence of EU-style risk-tier obligations re-
 2417 duces compliance burden but also reduces ex-ante clarity. The contemplated foundation-model reg-
 2418 ulation will determine whether the UK position consolidates around the principles-based approach or
 2419 moves closer to the EU AI Act’s prescriptive structure.

2420 **6.6.5 IP: UK Copyright Act §9(3) on Computer-Generated Works**

2421 UK copyright law contains a provision that, on its face, is favorable to AI-generated works in a
 2422 way that no other major jurisdiction’s copyright law is. Section 9(3) of the Copyright, Designs and
 2423 Patents Act 1988 provides that “[i]n the case of a literary, dramatic, musical or artistic work which is
 2424 computer-generated, the author shall be taken to be the person by whom the arrangements necessary
 2425 for the creation of the work are undertaken.” Section 178 defines “computer-generated” as generated
 2426 by computer in circumstances such that there is no human author of the work. The provision was
 2427 introduced in 1988 prospectively to address algorithmic-music and similar generative processes; it has
 2428 been little used in the intervening years and has not been authoritatively applied to modern generative
 2429 AI.

2430 The provision raises substantial interpretive questions in the AI context. The identification of
 2431 the “person by whom the arrangements necessary for the creation of the work are undertaken” could
 2432 plausibly fall on the user (who supplies the prompt), the provider (who operates the model), or the
 2433 model developer (who trained it); each candidate has been argued in commentary. The relationship
 2434 between §9(3) and the originality requirement under the *Infopaq*-derived “author’s own intellectual
 2435 creation” standard (still applied in UK law as a holdover from EU harmonization) is also contested:
 2436 if the standard requires human intellectual creation, §9(3) is in tension with it; if not, §9(3) supplies

2437 a coherent path to copyright in AI-generated work.

2438 The UK Intellectual Property Office's 2022 consultation on AI and intellectual property surfaced
2439 these questions and consulted on possible legislative responses, including repealing §9(3), amending
2440 it, or leaving it in place. The consultation concluded with the maintenance of §9(3) pending further
2441 development; subsequent UKIPO guidance has not significantly clarified its application. The 2024–
2442 2025 UKIPO consultation on copyright and AI training data (focusing on training-data uses rather
2443 than output authorship) deferred the output-authorship question.

2444 The result is that the UK is, in formal doctrinal terms, the jurisdiction most permissive of AI-
2445 generated work copyright, while in practice having the most uncertain regime, because the author-
2446 itative application of §9(3) to modern AI awaits litigation. The Beijing Internet Court's permissive
2447 application of ordinary copyright doctrine to AI-assisted work has, ironically, produced more con-
2448 crete jurisprudential guidance than the UK's purpose-built statutory provision.

2449 6.7 European Union

2450 6.7.1 Governance Paradigm: Comprehensive Ex Ante Regulation

2451 The EU governance posture toward AI is the most comprehensive ex-ante regulatory framework
2452 among the jurisdictions surveyed. The framework has five principal pillars: the AI Act (Regulation
2453 (EU) 2024/1689) addressing AI system providers and deployers; the Digital Services Act (Regulation
2454 (EU) 2022/2065) addressing intermediary services; the GDPR addressing personal data processing;
2455 the revised Product Liability Directive of 2024 extending strict-liability principles to software and AI;
2456 and the proposed AI Liability Directive, which at the time of writing has been withdrawn but whose
2457 substantive mechanisms (disclosure obligations, rebuttable causal presumptions) may be reintroduced
2458 in revised form or implemented through national law.

2459 The integrated framework is unique among the jurisdictions surveyed in two respects. First, it
2460 covers the AI value chain horizontally (model development, system integration, deployment, use)
2461 rather than only at particular nodes. Second, it integrates substantive obligations (the AI Act's risk-
2462 management requirements, the GDPR's data-protection requirements, the DSA's content-related obli-
2463 gations) with procedural mechanisms (conformity assessment, risk reporting, transparency, com-
2464 plaints handling) and with liability rules (the PLD's strict-liability extension, the prospect of AILD-
2465 style presumptions). The result is a regulatory architecture more demanding of providers than any
2466 other jurisdiction's but also more clearly specified.

2467 6.7.2 Criminal Law: Member State Competence

2468 The EU has no general competence over criminal law. Member states retain primary responsibility
2469 for substantive penal law and criminal procedure, subject to harmonization measures that the EU may
2470 adopt under TFEU Articles 82–83 in specified areas. The harmonization measures relevant to AI-
2471 extensible product harm are limited. Directive 2013/40/EU on attacks against information systems
2472 harmonizes minimum elements of computer-fraud and related offenses. Directive (EU) 2017/1371
2473 on the protection of the financial interests of the Union harmonizes serious-fraud offenses against EU
2474 financial interests. Directive 2011/93/EU on combating sexual abuse of children harmonizes related

2475 offenses including those involving AI-generated material.

2476 For the AI-extensible product configurations the article analyzes, the practical consequence is
 2477 that criminal liability is determined by member-state law, with the AI Act's administrative penalties
 2478 operating in parallel rather than in lieu. A multinational provider faces, in principle, the laws of each
 2479 member state in which its product produces harm or is operated; the Brussels Ibis Regulation governs
 2480 the jurisdictional allocation among member-state fora.

2481 The EU-level criminal-procedure mechanisms—the European Arrest Warrant, the European In-
 2482 vestigation Order, the European Public Prosecutor's Office for offenses against the EU's financial
 2483 interests—substantially facilitate cross-border enforcement within the EU but do not extend the sub-
 2484 stantive law itself. The AI-related criminal-law landscape within the EU is, accordingly, plural at
 2485 the substantive level and integrated at the procedural level—a combination distinctive among the
 2486 jurisdictions surveyed.

2487 **6.7.3 Tort: Product Liability Directive 2024 and AI Liability Directive**

2488 The 2024 revision of the Product Liability Directive (Directive (EU) 2024/2853, repealing Directive
 2489 85/374/EEC of 25 July 1985) materially extends EU product-liability law for AI-extensible product
 2490 analysis. The revised Directive expressly includes software (including AI systems) within the defini-
 2491 tion of product, treats absence of cybersecurity measures as a defect, addresses the development-risk
 2492 defense in narrowed terms, and introduces evidentiary mechanisms favorable to plaintiffs (a disclo-
 2493 sure obligation on defendants, a presumption of defectiveness in specified circumstances). Member-
 2494 state implementation is required by 9 December 2026. The implementation will substantially alter
 2495 the domestic tort landscape in each member state for AI-extensible product harm.

2496 The proposed AI Liability Directive (COM(2022) 496 final) would have provided an AI-specific
 2497 fault-based liability framework with two principal mechanisms: a disclosure obligation on AI providers
 2498 in respect of high-risk systems where the plaintiff has shown sufficient evidence of fault, and a re-
 2499 buttable presumption of causation between the disclosed fault and the harm. The AILD was the most
 2500 aggressive substantive law response to black-box causation among the AI-law instruments under
 2501 consideration in any jurisdiction. The Commission withdrew the proposal from its work programme
 2502 in February 2025; whether analogous mechanisms will be reintroduced through revised proposals,
 2503 integrated into national PLD implementations, or simply omitted is unsettled.

2504 The combined effect of the PLD revision and the AILD's uncertain status is that the EU position on
 2505 AI-related tort liability is substantially favorable to plaintiffs in the provider-defendant configuration,
 2506 with the multi-user joint-tortfeasance configuration that this article principally concerns receiving
 2507 less specific treatment. The German, French, and Italian implementations of the revised PLD will
 2508 produce the operative law for member-state actions; the harmonization that the Directive imposes
 2509 on the principal liability rules ensures substantial similarity, while leaving room for member-state
 2510 variation on unspecified points.

2511 The interaction with the AI Act deserves note. The AI Act's risk-management obligations operate
 2512 as substantive standards whose violation may, in tort proceedings, supply the fault element of fault-
 2513 based liability or, in PLD proceedings, support the defectiveness analysis. The combined effect is to
 2514 channel provider compliance behavior in directions that EU regulators have identified as appropriate,

2515 with corresponding ex-post liability consequences for non-compliance.

2516 **6.7.4 Administrative: AI Act as the Centerpiece**

2517 The EU AI Act (Regulation (EU) 2024/1689, the “Artificial Intelligence Act,” adopted 13 June 2024
2518 and entered into force 1 August 2024 with phased application) is the centerpiece of the EU AI regula-
2519 tory framework and the most extensive AI-specific administrative regulation among the jurisdictions
2520 surveyed.

2521 The Act adopts a risk-tier architecture distinguishing four levels. *Unacceptable-risk* practices
2522 (Art. 5: subliminal manipulation, exploitation of vulnerabilities, social scoring by public authori-
2523 ties, real-time remote biometric identification in publicly accessible spaces with limited exceptions,
2524 certain predictive-policing uses) are prohibited. *High-risk* systems (Art. 6 and Annexes I and III: sys-
2525 tems used as safety components of products covered by EU harmonization legislation, plus systems
2526 deployed in eight enumerated areas including education, employment, essential private and public
2527 services, law enforcement, migration, and administration of justice) face the Act’s principal obliga-
2528 tions: risk-management system (Art. 9), data governance (Art. 10), technical documentation (Art. 11),
2529 record-keeping (Art. 12), transparency and information provision to deployers (Art. 13), human over-
2530 sight (Art. 14), accuracy, robustness, and cybersecurity (Art. 15), conformity assessment (Arts. 43–
2531 44), registration in the EU database (Art. 71), and post-market monitoring (Art. 72). *Limited-risk*
2532 systems face transparency obligations (Art. 50: the user must be informed she is interacting with
2533 an AI system, AI-generated or manipulated content must be labeled). *Minimal-risk* systems face no
2534 specific obligations under the Act but are encouraged to follow voluntary codes of conduct (Art. 95).

2535 General-purpose AI models (Title VIII bis introduced in the final Parliament-Council compro-
2536 mise) face a separate set of obligations, including technical documentation, cooperation with na-
2537 tional competent authorities, copyright-compliance summaries, and—for GPAI models with systemic
2538 risk—additional model-evaluation, adversarial-testing, incident-reporting, and cybersecurity obliga-
2539 tions (Art. 55).

2540 Penalties are substantial. Violation of the prohibited-practice provisions can attract fines of up
2541 to EUR 35 million or 7% of global annual turnover, whichever is higher (Art. 99). Other violations
2542 attract lower but still substantial penalty maxima.

2543 The extraterritorial reach is broad (Art. 2): the Act applies to providers placing AI systems on the
2544 EU market regardless of establishment; to deployers located in the EU; and—most expansively—to
2545 providers and deployers located outside the EU where the output of the AI system is used in the EU.
2546 The provision echoes the GDPR Art. 3(2) and is expected to operate similarly in practice: providers
2547 globally adjust their conduct to AI Act requirements rather than operating dual systems for EU and
2548 non-EU markets.

2549 The Act interacts with sectoral regulation in complex ways. Sector-specific instruments (the med-
2550 ical device regulation, the machinery regulation, the financial services regulations, etc.) continue
2551 to apply alongside the AI Act, with explicit coordination provisions in the Act preserving sectoral
2552 competence. The DSA remains the principal instrument for content-related obligations of online in-
2553 termediaries; the AI Act’s transparency obligations on AI-generated content operate in parallel rather
2554 than in lieu. Application of the high-risk obligations from August 2026 and of the GPAI obligations

from August 2025 (with longer transitions for some elements) will produce the first full operational test of the architecture.

6.7.5 IP: DSM Directive Art. 4 TDM Exception

EU copyright law's distinctive provision for AI-extensible product analysis is the text-and-data-mining (TDM) exception of the Digital Single Market Directive (Directive (EU) 2019/790). Article 3 provides a non-overridable TDM exception for research organizations and cultural-heritage institutions for purposes of scientific research; Article 4 provides a broader TDM exception for any purpose, subject to the rightholder's express reservation in machine-readable form (the "opt-out"). The two-tier structure balances the EU's interest in supporting AI development against rightholder interests in controlling commercial training uses.

The opt-out provision under Art. 4 has been the principal point of contention. Effective implementation of the opt-out requires machine-readable signals at the level of individual works or datasets; the technical standards for such signals (robots.txt extensions, metadata standards, content-credentialing schemes) are in development but not yet uniformly adopted. The result is that the practical scope of permissible commercial AI training under Art. 4 depends on technical infrastructure that is itself contested.

The leading recent litigation engaging Art. 4 includes the Hamburg Regional Court's *Kneschke v. LAION* judgment of September 2024 and the GEMA litigation against OpenAI filed in November 2024 in Munich. The CJEU has not yet authoritatively addressed the application of Art. 3 and Art. 4 to AI training; the references that will reach the Court will substantially shape the operative position.

The EU AI Act's interaction with the DSM Directive is instructive. Article 53(1)(c) of the AI Act requires GPAI model providers to put in place a policy to comply with EU copyright law, including the Art. 4 opt-out; Article 53(1)(d) requires the publication of a sufficiently detailed summary of training-data content. The combination operationalizes the opt-out by requiring providers to take its operation into account in their compliance arrangements. Whether the operationalization will succeed in shifting the practical scope of permissible training is a question the next several years' enforcement will answer.

The output-authorship analysis follows the EU-level position discussed at §5.3: human creativity is required for copyright subsistence; mixed human-AI works are analyzed contribution-by-contribution under the *Infopaq/Painer* jurisprudence. The DSM Directive does not address output authorship.

6.8 Synthesizing the Comparison

The six jurisdictions surveyed cluster, on the dimensions most relevant to the article's argument, into three regulatory paradigms. The cluster boundaries are not sharp: each jurisdiction has features that fit awkwardly into the paradigm to which it is principally assigned, and the boundaries are themselves shifting as legislative and regulatory activity proceeds. The article nonetheless treats the clustering as a useful organizing device that supports a more parsimonious account of the comparative landscape than a six-way enumeration would permit.

Paradigm I: Continental Doctrinal Restraint (Germany, Japan). The defining features are (i) preservation of doctrinal coherence through resistance to bespoke AI doctrines, (ii) maintenance of criminal law's *ultima ratio* character through reluctance to extend criminal liability to AI-mediated conduct that does not satisfy the established constitutive elements, and (iii) reliance on civil and administrative law for the principal regulatory weight, with administrative regulation operating through soft law (Japan) or EU-derived hard law (Germany).

Paradigm II: Constitutional and Statutory Provider Protection (United States). The defining features are (i) constitutional constraints on government regulation of intermediary content, (ii) the broad statutory immunity of Section 230, (iii) the permissive regulatory posture toward AI development, and (iv) the corresponding under-protection of victims of distributed harm. The US position is, on the dimensions surveyed, the most sharply distinct from the others.

Paradigm III: Comprehensive Ex-Ante Regulation (EU, with the UK and China as variants). The defining features are (i) risk-graded ex-ante obligations for providers and deployers, (ii) integrated frameworks across data protection, AI, and intermediary-services regulation, and (iii) ex-post liability mechanisms (the revised PLD; the prospective AILD) calibrated to support the ex-ante framework. The UK adopts the duty-of-care approach for online services through the OSA but the principles-based approach for AI specifically, placing it between Paradigms I and III. China adopts comprehensive AI-specific regulation but with content-moderation orientation distinctive of the Chinese regulatory tradition; it shares Paradigm III's ex-ante orientation while diverging on the substantive content of regulation.

The three-paradigm taxonomy will be developed further in Chapter 7, which reorganizes the comparative material around specific issues rather than jurisdictions and articulates the trade-offs each paradigm represents.

7. Judicial Views and Cross-Jurisdictional Issue Analysis

7.1 Introductory Framing

Chapter 6 surveyed the six jurisdictions vertically: each jurisdiction's position was developed across criminal, tort, administrative, and IP doctrinal axes. This chapter reorganizes the same underlying material horizontally: it takes five specific issues that arise across the jurisdictions and examines the comparative pattern on each, highlighting points of convergence and divergence. The horizontal organization is more useful for practitioners and policymakers seeking to anticipate how a particular configuration will be treated across multiple fora; it is also more useful for the synthesis at §7.7, which articulates a typology of regulatory paradigms.

7.2 Issue 1: Indirect Perpetration Through AI Tools

The treatment of intentional users who employ AI-extensible products as tools to commit offenses is the area of greatest convergence across the six jurisdictions. Each reaches the conclusion that the intentional user is criminally liable for the constituent offense, through different doctrinal routes that share an underlying structural similarity.

Germany reaches the result through indirect perpetration (*mittelbare Täterschaft*, §25(1) StGB), with the AI treated as instrumentality (*Werkzeug*). The analysis follows the established doctrine for cases involving non-personal instrumentalities (printing press, firearm, computer); the AI character of the instrumentality does not, on the prevailing view, require doctrinal innovation.

Japan reaches the result through 間接正犯 (indirect perpetration) on the parallel structure, or alternatively through direct perpetration of the constituent offense (詐欺罪 for fraud) with the AI as the medium of the deceptive communication. The Japanese commentary (松宮孝明, 山口厚) treats the choice between these analytical routes as a matter of doctrinal preference; the result is the same.

China reaches the result through 间接正犯 (indirect perpetration) in penal-law theory, with cyber-crime applications increasingly common since the 2015 amendment. The Two Highs' interpretations on cyber-fraud have elaborated the application; the offense charged is typically the constituent offense (诈骗罪) rather than the 帮信罪 that is reserved for provider-secondary cases.

The United States reaches the result through ordinary principal liability for the constituent federal offense (typically wire fraud) and, where the user works with another, through aiding-and-abetting under 18 USC §2 as construed in *Rosemond*.

The United Kingdom reaches the result through ordinary principal liability for the relevant Fraud Act 2006 offense, with common-law accessorial liability (post-*Jogee*) available where multiple intentional users coordinate.

The European Union leaves the question to member-state competence, with the result that each member state's resolution follows the relevant national doctrine. The cross-border coordination mechanisms (EAW, EIO) facilitate enforcement but do not harmonize the substantive question.

The convergence is not coincidence. The intentional-user case is the configuration in which the moral and doctrinal stakes most clearly align: the user has chosen the harm, the AI is the means. Each jurisdiction's doctrinal apparatus, however calibrated, arrives at the conclusion that this configuration warrants criminal liability. The divergence—and the article's principal interest—arises downstream of intent.

7.3 Issue 2: Foreseeability Standards for Negligent Users

The treatment of negligent users—those whose mental state falls short of intent but whose conduct contributed causally to harm—is the area of greatest divergence among the surveyed jurisdictions on the substantive-doctrine axis.

Japan applies a doctrine of 具体的予見可能性 (concrete foreseeability) for negligent offenses where they exist, requiring that the defendant have been able to foresee the specific kind of harm that materialized. The standard is among the most demanding internationally and is rarely satisfied in AI-mediated configurations where the particular harm to the particular victim is, by hypothesis, unforeseeable. The combination of the demanding foreseeability standard with the absence of negligent variants for fraud and most property offenses produces near-zero criminal exposure for negligent users.

Germany applies an objective foreseeability standard calibrated to the professional role of the actor. The *erkennbares Risiko* analysis under negligence doctrine asks what a reasonable person in the

actor's professional position should have foreseen. AI professionals (developers, prompt-engineering specialists) face a materially higher standard than ordinary users; the calibration places Scenario β 's negligent user in a contested zone whose precise placement the BGH has not authoritatively addressed.

China operates with the 应当预见 (ought to foresee) standard for negligence, modulated by the policy considerations that animate Chinese prosecutorial discretion. The standard is in formal terms similar to the German, but its application is more variable; in AI-related cases the modulation tends toward conservative non-prosecution of the negligent user, reserving prosecutorial attention for the intentional user and for the provider through 帮信罪.

The United States applies a reasonable-person standard with a foreseeability filter, calibrated case-by-case through the common-law negligence framework. In the absence of comprehensive federal AI tort law, the standard's application varies by state and sector; the trend in the recent generative-AI defamation cases has been toward a relatively demanding standard for plaintiff-side foreseeability proofs.

The United Kingdom applies the reasonable-foreseeability standard of *Wagon Mound (No. 1)* [1961] AC 388 in tort, elaborated through the Caparo three-stage duty test for novel duty-of-care questions. The OSA's duty-of-care framework provides specific statutory standards for online services within scope.

The European Union, through the (currently withdrawn) proposed AILD, would have introduced an explicit rebuttable presumption regime that, while addressing causation rather than foreseeability, would in operation reduce plaintiff burdens in ways that effectively lower the foreseeability threshold. Without the AILD or analogous national legislation, the question is left to member-state law.

The same factual scenario—the negligent user of Scenario β , with the same prompts, the same publication, the same downstream misuse—could produce no criminal or civil liability in Japan, substantial civil liability in Germany or the UK, and (under a hypothetical AILD-style presumption) reduced-burden civil liability in the EU. The cross-jurisdictional divergence is substantial enough to make the choice of forum a substantial determinant of outcome, with the consequences for forum-shopping and rule-of-law equality the article identifies in §3.6.5.

7.4 Issue 3: Provider Liability and Safe Harbors

Provider liability and safe-harbor frameworks present the most striking cross-jurisdictional divergence the article identifies. The same provider, on the same conduct, can face materially different exposure across the six jurisdictions, with the practical consequence that provider design, deployment, and ToS choices are substantially shaped by the regulatory geography in which the provider operates.

Japan's *Winny* doctrine renders provider liability the narrow exceptional case: only providers whose subjective predominant purpose is unlawful use are exposed to complicity liability for users' offenses. The civil position parallels the criminal one for general tort liability; the notice-and-takedown framework imposes obligations on the providers within its scope but those obligations operate as duties of action upon notice rather than as bases for liability for underlying harm.

Germany's position combines the BGH's contributory-liability jurisprudence (broadly aligned with continental traditions of *Beihilfe*-based secondary liability) with the EU AI Act's ex-ante obligations as implemented in German law. The result is moderate provider exposure: criminal complicity is contested but not foreclosed; civil tort liability available where duty is established; administrative compliance burden substantial.

China's 帮信罪 plus the 2019 Interpretation make provider criminal exposure the most substantial among the surveyed jurisdictions, with the qualifier that prosecutorial discretion modulates actual exposure. The civil position under Civil Code Art. 1196 imposes joint-and-several liability on providers with constructive knowledge of users' infringement.

The United States's combination of Section 230 and the First Amendment makes the US the most protective of providers among the surveyed jurisdictions. The Section 230 question for AI-generated content is unsettled but the prevailing reading favors substantial provider protection. The civil position is similarly favorable, with ordinary tort doctrine providing the principal exposure but with the constitutional and statutory protections substantially limiting the operative scope.

The United Kingdom's OSA imposes statutory duties of care on in-scope providers, with substantial penalties for non-compliance. The combination of the OSA and ordinary common-law negligence produces moderate-to-substantial provider exposure, with the precise calibration depending on the OSA's scope determinations.

The European Union's AI Act imposes risk-graded ex-ante obligations across the AI value chain; the revised PLD imposes strict liability for defective AI-enabled products; the (uncertain) AILD would impose disclosure obligations and rebuttable presumptions. The combined effect is substantial provider exposure, principally through administrative penalties and civil claims, with the ex-post criminal exposure left to member-state law.

The divergence is consequential. A provider operating multinationally faces a materially different liability landscape in each market: minimal exposure in the US, moderate in Japan, substantial in the EU and the UK, potentially criminal in China. The result is that provider strategy must accommodate plural regulatory geographies rather than converge on a single global posture; the largest providers operate distinct compliance regimes for each market. The smaller providers without resources for multi-regime compliance face the strategic choice of operating in the most permissive jurisdiction (the US) and accepting limited international scope, or undertaking compliance burdens that may exceed their economic capacity.

7.5 Issue 4: Black-Box Causation in Adjudication

Courts in each surveyed jurisdiction have begun to engage with algorithmic causation in adjacent contexts, principally the algorithmic risk-assessment cases and the emerging generative-AI litigation. The engagement has been preliminary and no jurisdiction has yet articulated a clear doctrinal position for black-box causation in the AI-extensible product context. The courts' principal strategy has been avoidance: doctrinal questions that engage the underlying epistemic difficulty have been bracketed in favor of resolutions on adjacent grounds.

State v. Loomis, 881 N.W.2d 749 (Wis. 2016) addressed the use of the COMPAS risk-assessment

algorithm in criminal sentencing. The Wisconsin Supreme Court upheld the use against due-process challenge by emphasizing that the algorithm's output was one factor among many considered by the sentencing judge; the court explicitly declined to address the algorithm's internal operation, treating its proprietary character as a constraint on judicial inquiry. The implicit acceptance of black-box causation in the criminal-procedure context—the court's willingness to allow consideration of an output whose causal contribution to the sentencing decision was not itself transparent—has been read as both a precedent and a cautionary tale.

Ewert v. Canada, 2018 SCC 30, addressed the use of risk-assessment instruments by Canadian Correctional Service in respect of an Indigenous offender. The Supreme Court of Canada held that the Service had breached a statutory duty to ensure the accuracy of the information used in respect of the offender, particularly given the contested cross-cultural validity of the instruments. The decision is the most substantive recent common-law engagement with algorithmic causation in adjudication; its holding turns on a particular statutory duty and does not generalize to the broader question.

German cases on automated decision-making under GDPR Art. 22 have addressed the right to explanation in credit-scoring, insurance, and similar contexts. The *SCHUFA* judgment of the CJEU (Case C-634/21, judgment of 7 December 2023) clarified the application of Art. 22 to credit scoring and confirmed substantive rights of explanation, though the practical operationalization of the right against genuinely-black-box systems remains contested.

Chinese deepfake fraud enforcement has produced an accumulating body of CAC enforcement decisions and lower-court judgments, principally addressing intent and authentication rather than causation in the strict sense. The cases confirm provider exposure under the deepfake provisions but do not engage the black-box causation question directly.

Recent US generative-AI cases have begun to reach the underlying causation question in defamation, copyright, and employment-discrimination contexts. The pending litigation against generative-AI providers will, over the next several years, produce the first substantial body of US case law engaging the question.

The pattern is one of judicial caution: courts engage the algorithmic-causation question where the procedural posture forces them to and avoid it where they can. The article reads this pattern as an implicit acknowledgment of the difficulty's genuine intractability, and as a counsel for legislative rather than judicial response.

7.6 Issue 5: Cross-Border Enforcement Realities

The cross-border enforcement of AI-extensible product judgments encounters realities that shape the operative law in ways the substantive doctrines themselves do not fully capture. Several historical and recent episodes illustrate the pattern.

LICRA v. Yahoo! (Tribunal de grande instance de Paris, 20 November 2000) early established the principle that a national court could exercise jurisdiction over a foreign internet service on the basis of the service's effects within the territory and order the service to take measures to comply with national law. The subsequent US response (the Yahoo! court's refusal to enforce the French judgment, on First Amendment grounds in particular) illustrated the limits of effective enforcement

absent international cooperation. The pattern has recurred across many similar configurations.

Schrems II (Case C-311/18, judgment of 16 July 2020) addressed the transatlantic data-flow infrastructure and resulted in the invalidation of the Privacy Shield framework on grounds of inadequate US protection of EU personal data. The decision's practical effect—requiring providers to implement supplementary measures or to localize processing in the EU—illustrates the operative power of EU regulation to shape global provider conduct through extraterritorial enforcement.

The Budapest Convention on Cybercrime (Council of Europe, opened for signature 23 November 2001, entered into force 1 July 2004) provides the principal multilateral framework for mutual legal assistance in cybercrime cases among the more than sixty parties. Japan, Germany, the United Kingdom, the United States, Australia, and Brazil are parties; China and Korea are not. The Convention's procedural mechanisms substantially facilitate cross-border investigation and evidence gathering among parties; the gaps in coverage are felt principally in cases involving non-party jurisdictions.

The combined effect—traced in detail through Scenario ε at §3.6—is the “lowest-hanging-fruit” pattern that this article identifies as the central enforcement reality of AI-extensible product harm. Substantive doctrinal differences are partially neutralized by the convergence of practical enforcement on the most reachable defendants in the most cooperative jurisdictions. The neutralization is uneven: providers based in protective jurisdictions (the US) are largely insulated even where their conduct would attract liability elsewhere; individual users in restrictive jurisdictions (China, Brazil, certain EU member states) face substantial domestic exposure. The result is a pattern of enforcement asymmetry that the substantive doctrines do not address and that the rule-of-law commitments of each jurisdiction must engage on its own terms.

7.7 Three Regulatory Paradigms: A Synthesis

The five issues examined in the preceding sections collectively sustain the three-paradigm typology introduced at §6.8. This section develops the typology and articulates the principal trade-offs each paradigm presents.

Paradigm I: Continental Doctrinal Restraint (Germany, Japan). The paradigm preserves doctrinal coherence by resisting bespoke AI doctrines and maintains criminal law's *ultima ratio* character by reserving criminal sanction for clear intentional misuse. The cost is substantial gray zones: configurations of negligent contribution and innocent aggregation fall outside the criminal law's reach and are addressed, if at all, through civil and administrative channels of varying effectiveness. The benefit is the preservation of the protective principles that the restraint embodies: the privilege against self-incrimination, the in-dubio-pro-reo principle, the proportionality of punishment to fault. The paradigm is appropriate where the substantive harms are not so severe as to warrant the doctrinal expansion required to capture them comprehensively, and where civil and administrative mechanisms can adequately address the residual cases. The article takes the view that, on present evidence, this paradigm is defensible as a starting point and may need to be supplemented by ex-ante regulation to address distinctive AI risks.

Paradigm II: Constitutional and Statutory Provider Protection (United States). The paradigm maximizes provider freedom and minimizes regulatory intervention, on the substantive view that free

expression and innovation interests outweigh the costs of provider non-attribution. The paradigm relies on the market and on ordinary tort doctrine to address the residual cases. The principal cost is the difficult position of victims of distributed harm: the combination of Section 230 immunity, First-Amendment constraints, and the absence of comprehensive AI regulation leaves victims with substantially fewer remedies than in the other paradigms. The benefit is the avoidance of regulatory ossification and the support of the technological development that the US AI sector continues to dominate. The paradigm is appropriate where the substantive value of free expression and innovation is weighted heavily and the costs of provider attribution are substantial; the article takes the view that the paradigm's victim-protection deficit is its principal weakness and the principal focus of legitimate criticism.

Paradigm III: Comprehensive Ex-Ante Regulation (EU, with UK and China as variants). The paradigm front-loads regulatory burden through detailed ex-ante obligations and reduces the burden on ex-post adjudication. The principal cost is the substantial compliance burden, particularly on smaller providers, and the risk of regulatory ossification as the technology continues to evolve. The benefit is the predictability and consistency of regulatory expectations, the protection of victims through compliance-based risk reduction, and the explicit allocation of responsibility across the value chain. The paradigm is appropriate where the substantive harms are sufficiently severe and the regulatory capacity is sufficiently developed to support the architecture; the article takes the view that the EU AI Act's success or failure will substantially shape the comparative attractiveness of this paradigm over the next several years.

The three paradigms are not mutually exclusive in implementation. A jurisdiction can adopt elements of each paradigm in different doctrinal areas: the US could maintain Section 230 immunity for content while adopting EU-style ex-ante regulation for high-risk AI applications; the EU could supplement its ex-ante framework with US-style enforcement against intentional users; Japan could move from soft-law to harder regulation while preserving its 謙抑主義 criminal-law approach. The Colorado AI Act of 2024 and the contemplated UK AI Bill are explicit experiments in such hybridization. The article takes no position on the optimal hybrid; it observes that the three-paradigm typology captures the principal axes of variation, and that hybridization proceeds along those axes.

8. Policy Implications and Product Design Recommendations

8.1 The Humility of Diagnosis

The preceding chapters have identified structural dilemmas; they have not produced solutions. The recommendations developed in this chapter are accordingly preliminary and calibrated for incremental implementation rather than comprehensive reform. They are framed as principles and design choices that can be adopted by particular jurisdictions without requiring multilateral coordination, and that leave room for revision as the technology evolves and as practical experience accumulates.

The chapter's structure follows the natural pipeline from substantive law (legislative recommendations) through product architecture (design recommendations) to international cooperation. The recommendations operate at different levels of abstraction and are addressed to different audiences,

but they share a common substantive orientation: criminal law as *ultima ratio*; ex-ante regulation as the principal governance modality; clear safe-harbor frameworks as the bridge between regulatory burden and provider innovation; and product design as a substantive locus of risk management. The article does not claim that this orientation exhausts the available options or that it is uniquely correct; it claims that the orientation is consistent with the diagnostic analysis of the preceding chapters and supplies a defensible starting point for the policy work that AI-extensible product governance will require.

8.2 Legislative Recommendations

8.2.1 Maintain Criminal Law as *Ultima Ratio*

The article recommends against expanding criminal causation doctrines or evidentiary presumptions to accommodate AI-mediated harms. The recommendation has three specific elements.

First, jurisdictions should resist proposals to introduce burden-shifting devices akin to Japanese Penal Code Art. 207 (同時傷害特例) for AI-multi-user configurations. Such devices would, in operation, treat each contributor as solely responsible for the entirety of the harm in cases where the individual contribution cannot be established. The proportionality cost is severe: an actor whose contribution to a harmful output is, on any defensible measurement, modest is treated for purposes of punishment as if she had produced the entire harm. The procedural cost is comparable: the burden of proof on causation, fundamental to the protection of the innocent, is shifted to the defendant in a way that the original Japanese provision was deliberately confined to a narrow physical-injury context. The article reads the originally narrow scope as expressing a deliberate doctrinal judgment that the device should not be generalized.

Second, jurisdictions should resist proposals for strict criminal liability of AI providers. Strict liability in criminal law is generally disfavored in modern criminal-law systems; the exceptions are typically confined to public-welfare offenses with modest sanctions. The application of strict liability to AI providers, with the accompanying corporate-criminal sanctions that some proposals contemplate, would extend the exception in a way the principles do not support. The administrative-penalty regimes of the EU AI Act and analogous instruments are the appropriate vehicle for substantial financial sanctions on provider non-compliance.

Third, jurisdictions should reserve criminal sanction for clear intentional misuse and route negligent and ambiguous cases through civil and regulatory channels. The intentional-misuse configuration of Scenario α is criminally cognizable in every jurisdiction surveyed and should remain so; the negligent and mixed-culpability configurations of Scenario β and Scenario γ should generally be addressed through civil tort liability (with possible AILD-style presumptions on specific conditions) and through provider regulatory compliance. The administrative-channel preference is consistent with the broader trend toward ex-ante regulation as the principal governance modality.

The justification for this recommendation has three threads. Criminal law's commitment to *humility*—to refraining from expansion in domains where its evidentiary and substantive preconditions cannot be reliably met—is itself a doctrinal virtue. The historical record of expanded criminal liability under technological panic (Sieber's work on cybercrime history; the German legislative response

to the early hacking cases; the US response to the encryption “crypto wars” of the 1990s) illustrates the persistent risk of overreach. The proportionality principle, central to the substantive criminal law of every jurisdiction surveyed, constrains the kinds of attribution that criminal liability can support; the AI-multi-user configuration of Scenario γ is precisely the case in which the principle most sharply constrains.

8.2.2 Adopt Risk-Graded Ex-Ante Regulation

The article recommends that jurisdictions adopt risk-graded ex-ante regulation as the principal governance modality for AI-extensible products, taking the EU AI Act’s tiered approach as a template that can be adapted to local legal traditions. The recommendation rests on four substantive features that the AI Act’s architecture makes explicit.

First, *differentiated obligations by use-case risk*. The risk-tier structure permits regulatory burden to be calibrated to the substantive risk of the use case rather than imposed uniformly across all AI applications. Minimal-risk uses face minimal burden; high-risk uses face substantial burden. The calibration preserves regulatory capacity for the cases where it matters most while avoiding the over-regulation of low-stakes deployments. The specific risk-tier definitions of the AI Act are open to refinement in adaptation to other jurisdictions; the principle of risk-calibration is the core to be preserved.

Second, *compliance audits ex ante rather than litigation ex post*. The ex-ante compliance model concentrates regulatory effort on prevention rather than remediation, which is appropriate given the difficulty of remediation in cases where causation is opaque and harm is distributed. The ex-post liability mechanisms operate as backstops rather than primary tools.

Third, *safe harbors for compliant providers*. Providers who implement the prescribed compliance measures should benefit from clear immunity from secondary liability for the residual harms that compliance does not prevent. The immunity is the substantive quid for the compliance burden quo, and its absence in the current AI Act is, in the article’s view, a significant weakness that hybridization with EU PLD/AILD-style mechanisms should address.

Fourth, *mandatory transparency for high-risk applications*. Transparency obligations—about training data composition, model capabilities and limitations, decision processes for high-stakes uses—facilitate downstream accountability without themselves resolving the substantive questions. They support the operation of other doctrines (data-protection rights, civil tort analysis, administrative oversight) by reducing the information asymmetry that those doctrines presuppose.

The AI Act’s flaws are well-documented and the article does not recapitulate them in detail. The principal concerns are definitional ambiguity (the high-risk categories are sufficiently broad to invite interpretive disputes), compliance burden on smaller providers (the AI Act’s obligations are calibrated to provider sophistication that smaller entities may not possess), and the uncertain status of the AILD that would have completed the architecture. The refinements the article recommends include: clearer definitional guidance to reduce interpretive uncertainty; SME-specific compliance pathways with reduced burden; and re-introduction of AILD-style mechanisms (or member-state equivalents) to ensure that ex-ante compliance translates into ex-post liability protection for compliant providers and ex-post recovery for plaintiffs against non-compliant providers.

The recommendation is calibrated to operate in jurisdictions beyond the EU. The Colorado AI Act adopts substantially the same risk-tier architecture; the contemplated UK AI Bill is expected to incorporate analogous mechanisms; the Japanese soft-law guidelines could be hardened along the same lines without requiring radical restructuring. The article does not predict that all jurisdictions will converge on the EU model; it predicts and recommends that the risk-graded ex-ante approach will be the principal axis of convergence over the next decade.

8.2.3 Establish Provider Safe-Harbor Frameworks

The article recommends that jurisdictions establish AI-specific provider safe-harbor frameworks structured around clear, implementable, AI-specific safety measures. The recommendation is the natural complement to the ex-ante regulatory recommendation: where compliance with specified obligations is required, the incentive to comply must include clear protection against secondary liability for the residual harms that compliance does not prevent.

The substantive measures the framework should specify include: *content filtering* for known categories of unlawful or harmful output (CSAM, identified categories of fraud content, etc.); *abuse-detection systems* that monitor patterns of use indicative of misuse and trigger response measures (rate limits, suspension, reporting to authorities where legally required); *Terms of Service clarity* including specific prohibited-use catalogues, clear allocation of responsibility, and jurisdiction-specific provisions; *incident response* procedures with defined timelines for detection, mitigation, and notification; and *audit logs* sufficient to support post-incident investigation while respecting privacy constraints.

Provider implementation of the specified measures should trigger immunity from secondary liability for harms arising from user misuse. The immunity should be specific—tied to the categories of harm the measures are designed to prevent—rather than the general immunity Section 230 provides. The structure is analogous to the DMCA’s notice-and-takedown safe harbor, in which specified procedural compliance triggers protection from copyright-infringement liability.

The framework distinguishes itself from the US Section 230 position by conditioning immunity on substantive safety measures rather than providing it as a baseline. The Section 230 baseline provides no incentive for safety measures; providers can be ostentatiously irresponsible without losing protection. The proposed framework provides the protection but only for providers who have invested in safety. The framework distinguishes itself from the EU AI Act baseline by providing clean immunity for compliant providers; the AI Act imposes detailed obligations without specifying the corresponding immunity, leaving compliant providers exposed to the underlying liabilities (PLD, member-state tort) that compliance was meant to shield against.

The framework’s calibration must address several open questions. The granularity of the safety measures must be sufficient to provide compliance certainty without being so specific as to constrain legitimate innovation. The relationship between safe-harbor immunity and AILD-style rebuttable presumptions must be specified: presumptions should be available against non-compliant providers and unavailable against compliant ones, which produces the right incentives without overshooting. The immunity should not extend to providers whose conduct, despite nominal compliance, demonstrates bad faith (e.g., intentional facilitation of unlawful uses through ostensibly-compliant interfaces). These calibration questions should be worked out at the implementation level rather than at

the level of the basic framework.

8.3 Product Design Recommendations

8.3.1 Sandboxed Extensibility as a Design Norm

The article recommends, as a substantive design norm rather than a binding regulatory requirement, that AI-extensible products constrain their AI-mediated extension to defined operation sets or domain-specific languages, rather than offering open-ended code generation or arbitrary natural-language extension. The recommendation is addressed to product designers and to the regulatory bodies whose risk-assessment frameworks calibrate the permissible range of provider design choices.

The substantive case for sandboxing rests on four considerations. First, sandboxing reduces the space of possible harms *ex ante*. A product that permits only operations within a defined set cannot produce harms outside the set; the cumulative effect across many providers is a substantial reduction in the AI-related harm landscape. Second, sandboxing makes the system's behavior more predictable and auditable, supporting the broader transparency agenda that the regulatory frameworks have begun to pursue. Third, sandboxing provides a stronger basis for the "allowed risk" argument in doctrinal analysis. Both criminal and tort doctrine recognize that activities within an established class of permitted risks attract no liability for harms that materialize within the class; the more clearly defined the class, the more secure the doctrinal protection. Fourth, sandboxing eases satisfaction of regulatory transparency requirements; a product whose behavior is bounded by a defined operation set is easier to document, explain, and audit than one whose behavior is open-ended.

The cost of sandboxing is the reduced functionality available to users. The trade-off is genuine and the article does not claim that sandboxing is optimal in all configurations; it is more clearly defensible for high-stakes deployment contexts (healthcare, financial services, criminal justice) than for general consumer applications. The recommendation is that sandboxing be the default for high-stakes contexts and be available as an option for general contexts, with the choice transparent to users and calibrated to the substantive risks of the deployment.

8.3.2 Multi-Layer ToS Architecture

Provider Terms of Service should be structured to reflect the multi-doctrinal nature of the legal environment in which AI-extensible products operate. The article recommends three substantive features.

First, jurisdiction-specific provisions for the principal markets (US, EU, China, Japan, UK and significant secondary markets), with clear identification of which provisions apply to which users. Jurisdiction-uniform ToS that purport to apply identical terms globally typically fail under the unfair-contract-terms regimes of the more demanding jurisdictions and produce contractual confusion in those jurisdictions where the uniform terms are inapplicable.

Second, clear allocation of intellectual-property, liability, and data-protection responsibilities, calibrated to the doctrinal allocations the relevant jurisdictions actually permit. The three-layer copyright analysis of §5.3 establishes that some allocations the provider may seek (assigning all output rights to the provider, for example) cannot be effectively performed because no rights subsist in the AI-autonomous portion. The ToS should reflect this and avoid representations that may attract unfair-

trade-practices liability.

Third, specific provisions for the multi-user scenario. ToS that contemplate only single-user interactions miss the multi-user configurations that AI-extensible products increasingly support and on which the article’s central scenarios are built. Provisions addressing user awareness of shared-context features, the implications of those features for confidentiality and IP, and the allocation of responsibility for harms arising from context-aggregation effects, should be drafted explicitly.

8.3.3 Audit Logging Within Privacy Constraints

Providers should retain usage logs sufficient to support an audit trail in the event of dispute, while respecting the data-protection constraints that operate in their respective jurisdictions. The recommendation engages a genuine tension that no jurisdiction has resolved cleanly.

The substantive case for audit logging is straightforward. The black-box causation problem is partially mitigated where the provider has retained sufficient information about input prompts, context-aggregation events, and output generation to support post-hoc reconstruction of the causal chain to the extent the underlying technology permits. The reconstruction will not, in general, be sufficient for criminal-causation proof, but it can support civil-causation proof to the lower preponderance standard in many configurations.

The data-protection constraints operate in the opposite direction. GDPR Art. 5(1)(c) imposes data minimization; Art. 5(1)(b) imposes purpose limitation. PIPL Arts. 6 and 19 impose analogous constraints. APPI’s purpose-limitation framework is similar. Comprehensive logging of user prompts, context-aggregation events, and outputs may exceed what these provisions permit.

The principled trade-off the article recommends is calibrated logging keyed to risk tier. For minimal-risk deployments, minimal logging beyond what operational necessity requires; for high-risk deployments, comprehensive logging with retention periods calibrated to the likely period of dispute emergence; for intermediate cases, calibrated logging with sufficient detail to support audit while respecting the minimization principle. The specific calibration should be specified in the safe-harbor framework, with compliance triggering both data-protection defensibility (the logging is for a permitted purpose) and liability-allocation benefit (the audit trail supports post-incident analysis).

8.3.4 Prompt-Level Isolation

Providers should consider technical architectures that isolate user prompts to prevent the inter-user influence on which the mixed-culpability scenarios depend. The recommendation addresses the multi-user problem at its source rather than at the attribution stage.

The substantive choice is between shared-context architectures (in which prompts and outputs from recent users in a tenancy or topic cluster influence the model’s responses to subsequent users) and isolated-context architectures (in which each user’s interactions are processed in a context that excludes other users’ inputs). Shared-context architectures support several useful features (collaborative editing, organizational consistency, retrieval augmentation across team materials) and are widely deployed in enterprise products. Isolated-context architectures forgo these features but eliminate the aggregation effect on which Scenario γ ’s mixed-culpability problem depends.

The article does not recommend isolated-context as a universal default. The trade-off depends on the deployment context: a team-collaboration product may reasonably adopt shared context for the value-add it provides, while a consumer-facing product with high abuse-risk profile may reasonably adopt isolated context for the protection it provides. The recommendation is that the choice be made consciously, with explicit recognition of the multi-user implications, and that the choice be transparent to users (so that they can understand whether their inputs may influence other users' outputs and vice versa).

The technical implementation of prompt-level isolation has matured substantially in the past several years. Tenancy isolation, per-session context isolation, and selective retrieval augmentation that excludes other users' direct inputs are all available implementation patterns. The provider design choice is substantive rather than merely operational, with corresponding implications for the legal-risk profile of the deployment.

8.4 International Coordination Recommendations

The cross-border enforcement gap that Chapter 3 identified and Chapter 7 elaborated cannot be closed by domestic action alone. The article makes four brief recommendations on international coordination, each addressed to multilateral bodies and to the foreign ministries and justice ministries that conduct cross-border legal cooperation.

First, mutual legal assistance treaties (MLATs) should be updated for AI cases. Existing MLATs were drafted for paradigms in which evidence is geographically located and accessible through conventional production orders; AI-related evidence (model weights, context-aggregation logs, training-data provenance) presents distinctive technical challenges that the existing MLAT infrastructure addresses awkwardly. The Budapest Convention's 2022 Second Additional Protocol begins to address some of these issues for cybercrime cases generally; further work specifically on AI cases is warranted.

Second, a multilateral framework for AI provider notice-and-takedown could substantially reduce enforcement gaps in cross-border AI-mediated harm cases. The framework would build on the substantive coverage of national notice-and-takedown regimes (DMCA, EU DSA, UK OSA, China Civil Code Art. 1195–1196, Japan ISP-Liability Act) and provide procedural mechanisms for rapid cross-border notice and provider response.

Third, the rule-of-law implications of the “lowest-hanging-fruit” enforcement pattern should be explicitly recognized in international policy discussions. The pattern is not the result of deliberate design but the operative reality of a system that lacks multilateral allocation mechanisms; addressing it requires acknowledging that such allocation is needed.

Fourth, the Council of Europe Framework Convention on AI of 17 May 2024 (the “Vilnius Convention”) provides a starting point for multilateral cooperation on AI governance. The Convention's substantive provisions are at a high level of abstraction, but the institutional framework it establishes (meetings of parties, monitoring mechanisms, opportunities for elaboration of more specific obligations) supplies infrastructure that can support more concrete cooperation over time. The article recommends active engagement with the Convention's institutional machinery by all jurisdictions

3097 concerned with AI-extensible product governance, including those (the United States, China) that
3098 have not yet acceded.

3099 9. Conclusion and Open Questions

3100 9.1 What the Article Has Argued

3101 The article has advanced six interrelated propositions.

3102 First, AI-extensible products constitute a distinctive product paradigm whose legal analysis can-
3103 not be reduced to that of prior software paradigms. The paradigm's defining features—open-ended
3104 extension through natural-language prompts, processed by non-deterministic generative models that
3105 operate without reliable traceability—break the legal-doctrinal assumptions of traceable and deter-
3106 ministic user intervention on which prior doctrine relied.

3107 Second, the paradigm generates legal dilemmas that are *structural* rather than incidental. The
3108 dilemmas trace to features of the technology–practice complex itself rather than to defects of leg-
3109 islative drafting or judicial doctrine. Doctrinal strain that appears across multiple legal fields and
3110 multiple jurisdictions in homologous form is the diagnostic mark of structural rather than incidental
3111 difficulty.

3112 Third, three sources of dilemma—black-box causation, mixed culpability, cross-agent liability
3113 allocation—recur across the seven doctrinal fields surveyed. The recurrence is not coincidence but
3114 the consequence of the underlying paradigm features acting through different doctrinal vocabularies.
3115 The cross-doctrinal homology is the article's first theoretical contribution.

3116 Fourth, the six major jurisdictions surveyed cluster into three regulatory paradigms: continental
3117 doctrinal restraint (Germany, Japan), constitutional and statutory provider protection (United States),
3118 and comprehensive ex-ante regulation (EU, UK, China as variants). The clustering captures the prin-
3119 cipal axes of variation and supports a parsimonious account of the comparative landscape.

3120 Fifth, substantive doctrinal answers vary substantially across jurisdictions, but evidentiary and
3121 enforcement realities create *de facto* convergence on certain outcomes. Intentional users are reli-
3122 ably attributable across all jurisdictions; provider exposure varies dramatically; victim recovery de-
3123 pends substantially on factors orthogonal to the merits. The convergence-divergence pattern is itself
3124 a comparative-law phenomenon worth surfacing.

3125 Sixth, the article's contribution is diagnostic and cartographic rather than prescriptive. The dilem-
3126 mas may not admit of unified solution; legal systems may have to live with structural indeterminacy in
3127 this domain, as they have learned to live with similar indeterminacies in mass-tort causation and other
3128 unsettled fields. The article's positive proposals—ex-ante regulation as principal modality, criminal
3129 law as *ultima ratio*, safe harbors for compliant providers, sandboxed extensibility—are calibrated to
3130 operate in conditions of structural indeterminacy rather than to resolve it.

9.2 Methodological Reflection

The article's methodological choices warrant brief honest reflection. The constructed analytical case studies, defended at §3.1, are an instrument calibrated to the present moment in which actual case law is sparse and the technology is in active formation; their limits become more constraining as the case law accumulates and as the technology stabilizes. Subsequent work will have access to materials that this article cannot draw upon. The comparative analysis across six jurisdictions of substantially different legal traditions risks the standard comparative-law errors of false equivalence and hasty classification; the article has tried to mitigate the risks through anchor-case-driven analysis and explicit attention to the distinctive doctrinal architectures within which each jurisdiction operates, but cannot claim to have eliminated them. The decision to engage the problems prospectively, before substantial case law develops, is itself a methodological position that the article defends but that other scholars may legitimately contest. The cost of the article's breadth is real: each doctrine and each jurisdiction receives less depth than monograph-level treatment would allow, and specialists in each field will identify points where the article's coverage is too compressed to be satisfying.

The methodological positioning is, in a sense, itself an argument: diagnostic work has value even before solutions are available, and the early articulation of structural difficulties supports the collective work of legislative response, judicial development, and scholarly elaboration. The article's principal claim to contribution is not that it solves the problems it identifies but that it identifies them with sufficient clarity for that collective work to proceed.

9.3 Open Questions for Subsequent Scholarship

9.3.1 Doctrinal Refinement Questions

Several questions left open by the article's analysis call for further doctrinal work.

A complete formal model of layered attribution in mixed-culpability scenarios remains to be developed. The four-step framework sketched in §4.3 organizes the analysis but does not fully specify the doctrinal categories that should govern each step's application, the substantive content of the contribution threshold, or the relational categories that should govern the fourth step. Sustained doctrinal work in each of the major jurisdictions, building from the framework toward jurisdiction-specific articulations, would be valuable.

The specific implications of mechanistic interpretability research warrant continued attention. The article has taken the current state of the research as defining the practical scope of black-box causation; if interpretability research produces substantial advances in the next several years, the operative scope of the problem may shift. The legal implications of particular advances—feature-level decomposition of model processing, faithful attribution of output content to particular inputs—would warrant their own doctrinal articulation.

Sectoral analyses—how the dilemmas play out in healthcare, education, financial services, criminal justice deployment of AI-extensible tools—would substantially enrich the picture this article has drawn at the cross-doctrinal level. Each sector brings its own substantive standards and institutional features that interact with the structural sources in distinctive ways. The healthcare de-

ployments engage medical-malpractice and informed-consent frameworks; the criminal-justice deployments engage due-process and discrimination frameworks; the educational deployments engage academic-integrity and child-protection frameworks. The sectoral elaborations are partly underway in the existing literature but have not been integrated with the cross-sectoral analysis the present article attempts.

9.3.2 Comparative-Law Extension Questions

The comparative coverage of the article is selective and could be extended in several directions.

Civil-law jurisdictions outside Germany and Japan—France, Italy, Korea, the Netherlands, Spain—each present distinctive features that the present article has not engaged. The French *droit à l'image* and the Italian *diritto all'oblio* jurisprudence have particular bearing on AI-output configurations involving identifiable individuals; the Korean PIPL implementation and the Korean Constitutional Court's emerging digital-rights jurisprudence offer comparative material distinct from the Japanese position.

Common-law jurisdictions outside the United States and the United Kingdom—Canada, Australia, India, Singapore, South Africa—each have AI-related legal positions worth comparative engagement. The Canadian *Ewert* jurisprudence engaged briefly at §7.5 is the kind of material that deserves fuller treatment in extended comparative work. The Indian deepfake-regulation efforts of 2024–2025 supply distinctive material on the regulatory politics of AI in a major Global South jurisdiction.

Other major legal families—Islamic-law systems, Indic systems beyond Indian common law, Latin American jurisdictions with their own civil-law traditions—are entirely absent from the article's coverage. Extension to these systems would substantially enrich the comparative picture and would test the three-paradigm typology against legal traditions that may not fit cleanly within it.

The interaction with international human-rights frameworks (ICCPR Art. 19 on free expression and the related General Comment; ECHR Art. 10 and the Strasbourg jurisprudence on digital expression; the Inter-American Convention on Human Rights) is a further axis the article has not engaged. AI-related limits on expression, algorithmic-discrimination concerns under ICCPR Art. 26, and the developing right-to-explanation jurisprudence in regional human-rights systems each merit sustained treatment.

9.3.3 Technological Evolution Questions

The technological reference point of the article—the current generation of large language models deployed as AI-extensible products—will evolve, possibly in ways that render some specific claims obsolete even as the structural dilemmas persist. Four developments warrant flagging.

The transition from AI-as-tool to AI-as-agent paradigms further complicates the liability allocation. Where the AI is treated as agent rather than tool—autonomously selecting and executing sub-tasks in pursuit of broader user-specified goals—the attribution analysis must engage not just user prompts and AI processing but the AI system's own “decisions” in a stronger sense than the present article has needed to. The contemplated agentic-AI deployments in productivity, customer

service, and research assistance will press doctrinal categories that this article has held at the AI-tool level.

Multi-modal generative AI extends the harm landscape beyond text to image, audio, video, and combinations. Each modality presents distinctive harm profiles (deepfake video for fraud, AI-generated audio for impersonation, AI-image generation for CSAM and non-consensual intimate imagery) and engages the doctrinal categories somewhat differently. The comparative-law analysis the article has conducted at the text level has analogs at each modality, but the modality-specific elaborations are not straightforward.

Federated and on-device AI moves the processing locus to user devices and complicates the provider-attribution analysis. Where the model runs on the user's hardware and the provider's role is limited to model distribution, the analogies to traditional software-distribution liability become more apt; but the non-deterministic processing and aggregation effects that characterize centralized AI persist on-device. The intermediate configurations that combine on-device processing with provider-hosted updates and retrieval augmentation present distinctive doctrinal questions.

Advances in explainable AI may, over time, partially reduce the black-box causation problem. The present article has taken the problem as constitutive given the current state of mechanistic interpretability research; if the research succeeds in producing faithful per-output decompositions for production-scale models, the operative scope of the problem will shrink. The legal implications of partial success—faithful explanations available for some configurations but not others—will themselves require doctrinal articulation.

9.4 A Closing Note on Humility

Law has always lived with uncertain causation. The mass-tort jurisprudence that this article has used as a comparator was itself a generation's effort to come to terms with scientific causation that resisted clean attribution; the environmental-harm jurisprudence has worked through analogous problems for half a century. AI is, on the article's analysis, a sharp instance of an old difficulty rather than a categorically new problem. The doctrinal apparatus has resources for engaging it; the apparatus will need adaptation, but it does not need replacement.

Humility about what doctrine can resolve is itself a doctrinal virtue. The temptation to expand criminal liability, to invert evidentiary burdens, to extend regulatory reach beyond what the substantive analysis supports, is at its strongest in moments of technological anxiety. The article's principal practical recommendation—to resist these temptations and to address AI-extensible product harm through the more measured channels of ex-ante regulation, civil liability, and product design—rests on the conviction that law's protective principles are themselves substantive goods worth preserving even at the cost of the gaps that preservation produces.

The role of legal scholarship in technology governance is, at its best, a slow patient work of clarification rather than a race to prescribe. The article hopes to have contributed to that work in respect of the AI-extensible product paradigm; whether the contribution proves useful will be measured by the work it supports rather than by the answers it provides.