



**EUROPEAN MASTER
IN LAW & ECONOMICS**

Master's Thesis

European Master in Law and Economics (EMLE)

University of Arizona

**Data Holding Rights and Fair Use Defense:
A Welfare-Based Analysis of Chinese Policy**

Ziqing Wang

ziqing.wang@emle.eu

10.08.2025

Authorship Declaration

I hereby declare and confirm that this thesis is entirely the result of my own work except where otherwise indicated. I acknowledge the supervision and guidance I have received from Prof. Simone M. Sepe. This thesis is not used as part of any other examination and has not yet been published.

10.08.2025

Ziqing Wang

Ziqing Wang

Abstract

Based on China's "three-rights separation" data rights framework, this paper constructs a two-stage complete information game model to assess how conferring data holding rights affects investment incentives and overall social welfare, and to identify the institutional conditions under which fair use should be permitted. The study finds that while property rights protection can correct the market failure of "zero investment," systemic underinvestment remains because data holders cannot appropriate innovation spillovers. To address this gap, the paper proposes a welfare criterion, $\psi = (\beta + \gamma - \delta) \alpha - \varepsilon$, providing a quantifiable standard that links measurable parameters to fair use determinations and differentiated policy design. This study further suggests that different types of data should be subject to differentiated policies and introduces a fair use system as a structural regulatory mechanism to dynamically balance investment incentives and data circulation.

JEL-codes: K11, K20, C7

Keywords: Data Ownership, Fair Use, Data Use Regulation

Table of Contents

1. INTRODUCTION.....	1
2. LITERATURE REVIEW	5
2.1 THE DATA ENTITLEMENT DEBATE	5
2.2 LEGAL CONSTRUCTION OF DATA PROPERTY RIGHTS.....	6
2.3 RIGHTS LIMITATIONS IN DATA GOVERNANCE	7
3. BASELINE MODEL	9
3.1. THEORETICAL FOUNDATION	9
3.2. GAME STRUCTURE	10
3.2.1 <i>Players, Time Series, and Information Structure.....</i>	<i>10</i>
3.2.2 <i>Investment, Cost, and Social Welfare Function</i>	<i>11</i>
3.2.3 <i>Equilibrium Analysis and Core Findings.....</i>	<i>14</i>
3.3. MARGINAL EFFECTS OF PARAMETERS AND POLICY IMPLICATIONS.....	16
3.4. DATA TYPES AND DIFFERENTIATED POLICY	17
3.5 ROBUSTNESS TESTS UNDER INCOMPLETE INFORMATION	20
4. WELFARE-BASED ASSESSMENT OF FAIR USE DEFENSE.....	23
4.1 FROM PRINCIPLE TO EXCEPTION	23
4.2 THEORETICAL CONSTRUCTION OF FAIR USE DEFENSE.....	24
4.2.1 <i>Basic assumptions</i>	<i>24</i>
4.2.2 <i>A Welfare-Based Derivation of the Fair Use Defense.....</i>	<i>25</i>
4.3 PARAMETER IDENTIFICATION AND CLASSIFICATION REGULATION.....	26
4.3.1 <i>Observable behavioral parameters</i>	<i>26</i>
4.3.2 <i>Fair use analysis based on data type and purpose.....</i>	<i>29</i>
4.4 EMPIRICAL APPLICATION: DEEPSEEK DISTILLATION CASE	33

4.4.1	<i>Case Overview and Theoretical Positioning</i>	33
4.4.2	<i>Parameter Identification and Quantitative Analysis</i>	33
4.4.3	<i>Fair use judgement</i>	35
5	POLICY IMPLICATIONS AND CONCLUSION	36
5.1	POLICY IMPLICATIONS	36
5.1.1	<i>Ex ante regulation: data classification and differentiated enforcement</i>	36
5.1.2	<i>Ex post liability: establishing fair use defense</i>	38
5.2	CONCLUSION	40
6	REFERENCES	43
7	APPENDIX	46
7.1	COMPLETE LIST OF MODEL PARAMETERS	46
7.2	TECHNICAL DESCRIPTION OF USER B	46
7.3	FORMALIZATION OF CORE ASSUMPTIONS	47
7.4	DETAILED DERIVATION OF PROOF	47

1. Introduction

In early 2025, OpenAI alleged that DeepSeek used billions of GPT-4 API calls to “distil” a competing model without authorization. OpenAI argued that such large-scale extraction effectively appropriated the returns on its substantial model-training investment. However, some reports noted that DeepSeek’s model operated at a fraction of GPT-4’s price and could lower access costs for domestic developers. The dispute illustrates a structural tension in data governance. If all secondary uses are freely permitted, original investors may reduce or withdraw investment in costly data curation, fearing free-riding problem. Conversely, if all such uses are prohibited, downstream innovation—particularly in AI, which relies heavily on access to high-quality data—may be stifled and market power concentrated in a few incumbents. This trade-off lies at the core of debates over whether, and under what conditions, a fair use framework should temper exclusive control to reconcile investment incentives with the broader social value of data reuse.

To mitigate this conflict, different jurisdictions have adopted different approaches. All regimes seek to internalize data’s network externalities because the risks of underinvestment and excessive deterrence can undermine innovation. U.S. courts provide limited scraping exceptions grounded in competition policy (e.g. *hiQ Labs v. LinkedIn*), whereas the European Union relies on a sui-generis database right, the portability provisions of the GDPR, and, more recently, the EU Data Act. China’s December 2022 Opinions on Building a Basic Data System (hereinafter the “Data Twenty Measures”) responds by proposing a “three-rights separation” that states a new form of data entitlements to reconcile the trade-off.

The Data Twenty Measures divides data entitlements into three rights. First, the data holding right is legally entitled to data holders who organize the collection and storage of raw data.

It grants *de facto* control over servers and raw datasets, together with a defensive claim like quasi-possession in civil law. Second, the data processing and usage right authorizes acts of cleaning, transforming, aggregating, or analyzing lawfully held data. Crucially, this entitlement is designed to be splittable, signaling legislative intent to create a modular market for data use. Third, the data product management right governs the commercial development of derivative outputs.

The data holding right granted to data holders is conceptualized as a usufructuary right (a right *in rem*) rather than full ownership. The “three-rights separation” design draws inspiration from China’s rural land reform, when the separation of ownership, contracting rights, and management rights released the production potential of idle plots of land. Similarly, separating use from possession in the digital realm is expected to encourage data reuse without stripping holders of baseline control.

Entitling data holders such limited *in rem* right is grounded in Locke’s labor principles and utilitarian incentive theory. Following Locke, the party that mixes labor and capital with unowned resources will acquire a lawful property entitlement. Data holders invest in cleaning, labelling, and security compliance, so excluding free riders protects the fruits of that effort. From a utilitarian perspective, clearly defined rights prevent a digital “tragedy of the commons”. By ensuring that investors can appropriate part of the surplus they create, the law incentivizes continual data collection.

Rights fragmentation theory further provides legitimate justification of conferring data holding rights rather than full ownership. The goal is to carve out a tradable element from an indivisible resource bundle so that the resource can circulate and be traded while preserving core control. If only an undivided ownership exists, the data holder faces a rigid binary choice—either transfer the whole asset or refuse any access at all—an outcome that

constrains the flow of production factors whose value depends on large-scale recombination. By splitting the entitlement, the legislature allocates residual rights to the original collector, safeguarding ultimate control, yet creates a tradable processing-and-use right. In short, segmentation transforms a non-marketable, all-or-nothing claim into a marketable asset class and thereby maximises allocative efficiency without surrendering systemic safeguards.

While data holders enjoy autonomous control, their usufructuary entitlement cannot be boundless. Excessive fragmentation of exclusive rights raises negotiation costs and triggers the tragedy of anti-commons. To prevent this property regime becoming a barrier to innovation, the legal framework must incorporate limitation. Because data, like intellectual property, is infinitely replicable and non-rivalrous, overly strong exclusivity has a chilling effect on follow-on innovation and technological convergence. Drawing on copyright legal practice, a fair use defense provides a proven template. It authorizes narrow, socially beneficial uses without excessively eroding the returns of original investors.

Optimal rights allocation requires balancing collecting costs, data users' value creation potential, competitive effects, and overall welfare implications. While China's policy formally creates distinct categories, its core objective is to strike a balance between data holder protection and data user access.

This raises two research questions: (1) how granting data holders holding rights affects their investment decisions and overall social welfare; (2) under what conditions institutional should designs allocate processing and usage rights to data users? To answer these questions, this paper continues Gallini and Winter's (1985) dynamic analytical framework for the "investment-infringement" structure, employing a game theory model to identify the dynamic linkage between property rights structure and investment incentives. Furthermore, it analyzes the institutional boundaries of "fair use" through comparative marginal social welfare.

The following chapters are structured as follows: Chapter 2 reviews existing literature on the establishment and restriction mechanisms of data property rights in both Chinese and global contexts; Chapter 3 constructs a baseline model to analyze the incentive effect and social welfare structure under pure property rights protection; Chapter 4 introduces the fair use defense and proposes welfare comparison conditions and an operational identification framework; Chapter 5 summarizes the policy implications and conclusion.

2. Literature Review

2.1 The Data Entitlement Debate

The data-entitlement debate in China divides scholarship into property-rights advocates and behavioural-regulation supporters.

Property-rights scholars argue that gaps in contract, IP and competition law, combined with free-riding and high bargaining costs, leave core investment incentives inadequately protected (Shen, 2020; Long, 2017; Cheng, 2018; Wang, 2023)¹. Without entitlements, data markets suffer an under-investment trap: collection is costly yet produces large spillovers that collectors cannot fully appropriate (Xu, 2018). As a result, data allocation can devolve into an uncoordinated scramble that discourages both investment and sharing (Shen, 2020). Besides, markets need trust to function, but uncertainty about data rights undermines confidence (Shen & Liu, 2022) and raises transaction costs (Wang et al., 2022). Such coordination and trust problems mirror Farrell's (1987) information paradox, in which valuable information may fail to be traded because it must be revealed before its value can be assessed, undermining incentives to invest.

Behavioural-regulation scholars advance three objections. First, they doubt that stronger rights are even needed: firms already generate data as a by-product of routine operations² (Fu, 2022; Dai, 2021); technical locks and rapid obsolescence deter free riding (Jin, 2020); and extra exclusion could impose deadweight loss (Wang, 2019). Second, proliferating entitlements risk an anticommons problem, where unclear and overlapping claims leave

¹ Contract claims fail once datasets require multi-party consent; copyright and trade-secret rules exclude non-innovative or publicly observable data; competition-law standards remain too indeterminate to offer ex ante certainty, increasing bargaining costs.

² Customer-service logs, click-streams and sensor readings arise automatically in daily operations; firms collect them for internal optimisation, not for standalone sale, so added legal rents may over-reward inevitable activity.

resources idle (Heller 1998; Stepanov 2020; Hu 2022). Third, current approaches work reasonably well. Companies protect data through possession, contracts, and secrecy (Stepanov, 2020) without apparent market failure (Jin, 2020). For example, the EU's data portability rights create quasi-property effects through competition law, suggesting formal property rights may not be necessary (Graef, 2016).

2.2 Legal Construction of Data Property Rights

China's Data Twenty Measures structures data rights by separating data entitlements into three rights. This approach transforms a complex bundle of interests into standardised, tradeable units, much like what Merrill & Smith (2000) describe as “optimal standardization” in property law, where limiting rights to a finite set of standard forms reduces information and transaction costs. This middle path departs from the ownership-vs-access binary by treating property as a set of relationships (Hohfeld, 1913) and by embedding power dynamics in information flows (Cohen 2019).

Two theoretical frameworks guide this translation. Labor theory, following Locke (1964), suggests that mixing labor with resources creates ownership. Data processors invest effort in collection and curation, justifying their claims (Long, 2017; Qian, 2021). Bundle of rights theory, rather than focusing on control over objects, it examines relationships between people regarding resources (Xu, 2021). This perspective, despite Penner's (1996) critique that it lacks coherence, helps explain data's complexity—multiple parties, multiple uses, multiple values (Wang, 2022). It also underscores the need to align rights design with behavioural realities. For example, the so-called *privacy paradox* shows that individuals often disclose personal data despite professing strong privacy concerns—not due to genuine indifference, but because of imperfect heuristics and cognitive constraints (Solove, 2021). Recognising

such behavioural patterns supports flexible “dynamic limitation mechanisms” that adjust specific powers without dismantling the overall framework (Bao & Xiong, 2020). Some propose “rights blocks”³—pre-packaged combinations of rights and obligations tailored to specific uses (Xu, 2023).

2.3 Rights Limitations in Data Governance

Rights in data cannot be defined solely by legal doctrine; economic forces ultimately shape how they operate in practice. Because data are non-rival yet often valuable only when access is limited, simple property rules misfire. Coasean bargaining collapses once transaction costs arise and exclusivity affects aggregate welfare (Bergemann & Bonatti, 2019). Information asymmetry produces the Farrell paradox, highlighting the need to balance holder control with user access (Farrell, 1987; Jones & Tonetti, 2020). Verification gaps—such as the inability to monitor how data are collected, used, or valued—create moral hazard (Dosis & Sand-Zantman 2023), and over-collection can generate negative externalities when incentives outrun social need (Acemoglu et al., 2022). Both overly strong and overly weak rights can backfire. Credit availability falls when borrowers control data too tightly (He et al., 2020), while perfect exclusion is impossible and often undesirable (Acquisti et al., 2016).

Economic analysis reveals why absolute property rights fail for data governance and justifies the introduction of rights limitations to calibrate exclusivity. Tirole (2023) argued that when data becomes infrastructure for downstream innovation, mandatory access obligations may be necessary. This recognition drives specific limitations of exclusive rights.

³ For instance, a “rights block” might bundle user browsing data access with specific advertising uses, time limits, and user revocation rights—creating a standardized, tradeable package.

Copyright law offers a useful starting point, but data differ in keyways: they are non-rival, gain value through aggregation, and depend on network effects. As a result, traditional fair-use factors—purpose, nature, amount, and market effect—require adaptation rather than wholesale replacement. The EU’s portability mandate shows how this can work in practice: users can retrieve their own records, opening the door to market entry without changing formal ownership (Graef, 2016). Patent-law FRAND rules offer another analogy: regulated, non-discriminatory access can prevent holdup while preserving rents (Lemley & Shapiro, 2007; Contreras, 2019). In general, digital economics supports graduated, context-specific entitlements over a rigid property/no-property split (Goldfarb & Tucker, 2019). This approach strengthens the case for a fair use defense in data governance—one that can be applied and tested across different data types and usage contexts.

3. Baseline Model

3.1. Theoretical foundation

The theoretical framework of this study is based on the "investment-infringement" game theory model of Gallini and Winter (1985), in which exclusive rights allow innovators to recoup R&D costs, and equilibrium investment depends on detection probability and compensation intensity. Their three-stage structure—investment, licensing, and potential infringement—captures how later licensing terms feed back into earlier investment incentives.

The temporal structure of the framework is consistent with the research process of data governance issues. Data collection, like patent research and development, requires upfront sunk costs, and its output is also a non-competitive, easily replicable intangible asset. Therefore, investors face homogeneous "unauthorized use" risks afterwards. At the same time, the model is highly compatible with the policy tools of "establishing property rights - strengthening law enforcement - setting reasonable use exceptions", which echoes the three-stage governance idea advocated by the "Twenty Data Measures". By focusing initially on complete information settings, we can establish the core mechanisms cleanly before introducing real-world complications. While information asymmetries and enforcement uncertainties certainly matter in Section 3.5, this baseline model generates robust insights about optimal property rights design that hold even in more complex environments.

Following Gallini & Winter, based on the framework, this paper adjusts the characteristics of the data market: the main structure of the model is simplified from the duopoly competition in the patent market to the "one-to-many" relationship in the data market. We adopt a probabilistic expected compensation mechanism to characterize the realistic characteristics

of data law enforcement. The model retains the core mechanism of "exclusive rights promote investment" and reflects the essential characteristics of the data economy: value creation mainly relies on scale and network effects, rather than technological monopoly. The game model constructed in this chapter is used to analyze the economic effects of data holding rights under the "three-rights separation" framework of the "Twenty Data Measures". The core research question is how giving data collectors clear resource holding rights affects their investment decisions and total social welfare?

3.2. Game structure

3.2.1 Players, Time Series, and Information Structure

To focus on the pure effect of property rights establishment on investment incentives, this paper first constructs a two-stage⁴ dynamic game. The participants include a data resource holder, H , and a representative data user, B . This simplified structure captures the fundamental tension between investment incentives and usage efficiency. H represents the enterprise that holds data rights under the "Twenty Data Measures" and must decide its level of investment in data collection in the first stage. B represents the aggregate effect of potential data users, who, after observing the data's value, decide whether to use it without permission in the second stage.

Furthermore, the perfect information assumption is adopted⁵: all parameters are common knowledge, and B can accurately observe the data's value. While this simplifies away the information asymmetries prevalent in real data markets, it reflects an increasingly common

⁴ The model adopts a simplified two-stage structure and does not consider the possibility of market transactions. This simplification is based on two practical considerations: first, serious information asymmetry is prevalent in data markets; second, data pricing faces fundamental difficulties. These factors together lead to excessively high transaction costs.

⁵ The perfect information assumption is intended to focus on the incentive effects of property rights allocation and facilitate identification of the marginal impact of property rights strength itself. Regulators and courts may face multiple information constraints.

scenario: major platforms like AWS and Alibaba Cloud now provide comprehensive data samples and quality metrics, enabling informed assessment before purchase. This transparency trend makes our assumption reasonable for analyzing core incentive mechanisms, though Section 3.5 explores how results change when information is incomplete or enforcement uncertain.

The time series structure of the game is as follows: In the first stage, H , uncertain about B 's future usage decisions, chooses an investment level to maximize expected returns. Investment determines the scale and quality of the data, which in turn influences its economic value. In the second stage, after observing the realized data value, B chooses between "using without permission" and "abandoning use"⁶. The user's decision depends not only on the expected net benefits of use, but also on the probability of intervention by regulators or courts.

3.2.2 Investment, Cost, and Social Welfare Function

The key decision variable for data holders is the level of data investment x ($x \geq 0$), which determines the scale and quality of the data. This investment generates basic value αx , where the conversion coefficient $\alpha > 0$ reflects data processing capabilities—collection technology, algorithmic sophistication, and governance systems. For instance, Google's Chief Economist Hal Varian (2019) documents how data value in advertising markets scales proportionally with collection volume, while Alibaba reports similar linear patterns in e-commerce

⁶ This model does not include "users pay for authorization" in the strategy set, nor does it include the intermediate stage where "owners decide to share or block." The reason is that if owners are willing to share, both parties can internalize the externality through free contractual transactions, resulting in zero marginal effect on property rights allocation. This paper focuses on the other extreme—the distortions in investment incentives and social welfare that arise when owners refuse to share due to confidentiality or competitive considerations. To highlight the research focus and maintain model simplicity, the original three-stage game of "Stage 1: Investment $\rightarrow$ Stage 2: Sharing/Blocking $\rightarrow$ Stage 3: Infringement" is condensed into two stages: "Investment $\rightarrow$ Infringement." The sharing option is deemed to be resolved through contractual mechanisms and is therefore omitted from the analysis without affecting the evaluation of the effectiveness of the "Establish Ownership Rights-Fair Use" policy combination.

applications. A company with superior algorithms extracts more value from identical data resources, captured by a higher α .

This basic value has two positive effects. First, from the perspective of private value, users can extract private benefits $\beta\alpha x$ from it. Where $\beta \in (0,1)$ represents the proportion of private value to basic value. β reflects the private appropriability of data value. Some data applications (such as precision marketing) are easy to convert into private benefits, while others (such as basic research) are difficult to internalize their value. Second, data use also generates knowledge spillovers $\gamma\alpha x$, or innovation externalities. Where $\gamma > 0$ reflects the strength of this knowledge spillover effect. If we further break it down, γ can be divided into two parts: one is direct spillover γ_{direct} ; the other is the sum of indirect dissemination $\gamma_{indirect} \cdot \theta$, that is, $\gamma = \gamma_{direct} + \gamma_{indirect} \cdot \theta$. Therefore, the total positive benefit generated by data use is:

$$V(x) = (\beta + \gamma)\alpha x \quad (1)$$

When $\beta + \gamma > 1$, the total social value of data usage exceeds its initial base value. However, $\beta + \gamma$ can also be less than 1, indicating that this part of the data may have value dissipation or realization costs.

The second dimension is cost. The investment cost function adopts the quadratic form:

$$C(x) = cx^2 \quad (2)$$

This quadratic specification assumes increasing marginal costs of data investment that matches industry realities. IDC (2023) reports that expanding data infrastructure from petabyte to exabyte scale increases per-unit costs by roughly 40%, while Microsoft Azure's pricing tiers exhibit similar patterns of rising marginal costs. The increasing costs reflect mounting technical bottlenecks, management complexity, and quality control challenges as

data operations scale. For example, expanding from 1 million users to 10 million requires not only more servers and bandwidth but also a more complex data governance system.

Data use also generates two types of negative externalities. Competitive harms are:

$$L_{(competition)(x)} = \delta \alpha x \quad (3)$$

where $\delta \in [0,1]$ can be decomposed into three dimensions:

$$\delta = \delta_{market} + \delta_{bargaining} + \delta_{innovation} \quad (4)$$

These three dimensions describe the different types of damage caused by data leaks to enterprises: market share transfer effect (direct customer loss due to data leaks), bargaining power weakening effect (loss of bargaining advantage after suppliers and customers know internal data), and innovation advantage loss effect (loss of first-mover advantage due to imitation by competitors) :

The privacy cost is:

$$E_{privacy(x)} = \varepsilon x \quad (5)$$

($\varepsilon \geq 0$, note that privacy costs are related to the amount of data rather than its value.)

This setting reflects the cumulative nature of privacy risks—the larger the amount of data, the higher the potential risk of privacy leakage, and is not linearly related to the economic value of the data.

In terms of institutional parameters, the probability of regulatory intervention $q \in (0,1)$ reflects the differentiated responses of regulators and courts to different types of data use. The underlying implication of this parameter is that regulators are conducting an implicit cost-benefit analysis. For uses with high innovation externalities or low competitive harm, regulators may choose not to intervene, even if such uses are technically detectable. The compensation multiplier $d > 0$ determines the intensity of the penalty ($d=1$ for compensatory

damages, $d > 1$ for punitive damages). The product of the two, qd , determines the expected cost rate of unauthorized use, which is a key variable influencing user decisions.

Combining the above factors, the social welfare function is:

$$SW(x) = [(\beta + \gamma - \delta)\alpha - \varepsilon]x - cx^2 \quad (6)$$

Defining the net social value coefficient $\psi = (\beta + \gamma - \delta)\alpha - \varepsilon$, then $SW = \psi x - cx^2$. This concise expression contains rich economic implications: when $\psi > 0$, data use generates net social value, and the optimal social choice is to encourage data collection and use; when $\psi \leq 0$, negative externalities outweigh positive value, and restricting data activities may be more reasonable.

3.2.3 Equilibrium Analysis and Core Findings

Using backward induction to solve this dynamic game, we begin our analysis by examining the usage decision in the second stage. Data user B, as a rational economic agent, only receives benefits that can be internalized. When B chooses "unauthorized use," his payoff is $\pi_B = (1 - qd)\beta\alpha x - \varepsilon'$ (where $\varepsilon' > 0$ represents a minimal cost of use); when he chooses "abandoning use," $\pi_B = 0$. Therefore, the usage decision follows a simple rule: when $\beta > 0$ and x is large enough such that $(1 - qd)\beta\alpha x > \varepsilon'$, B chooses to use; when $\beta = 0$ or $x = 0$, B chooses not to use.

Based on the equilibrium results of the second stage, we analyze the investment decisions of the first stage. Here, the role of the property rights system begins to emerge.

Proposition 1 (Market Failure without Property Rights)

In the absence of property rights protection ($d=0$), the equilibrium investment level $x_0^*=0$ and social welfare $SW_0=0$.

When $d = 0$, the objective function of the holder H is $\pi_H = -cx^2 - \delta\alpha x$. Since $\pi_H < 0$ for any $x > 0$, a rational holder will choose $x_0^* = 0$.

While this result is simple, it reveals the fundamental role of property rights: the absence of property rights protection leads to complete market failure, leaving data holders facing a dilemma of "only costs and no benefits." This explains why many valuable data resources were not developed in the early days of the digital economy.

Proposition 2 (Investment Incentives under Property Rights Protection)

Under property rights protection ($d > 0$), the equilibrium investment level is:

$$x_1^* = \frac{(qd\beta - \delta)\alpha}{2c} \quad (6)$$

The objective function of owner H is $\pi_H = (qd\beta - \delta)\alpha x - cx^2$. The first-order condition $\frac{\partial \pi_H}{\partial x} = (qd\beta - \delta)\alpha - 2cx = 0$ yields the above result. The second-order condition $\frac{\partial^2 \pi_H}{\partial x^2} = -2c < 0$ confirms the maximum. When $qd\beta \leq \delta$, $\partial \pi_H / \partial x \leq 0$ holds for all $x > 0$, so $x_1^* = 0$.

This proposition reveals the core mechanism of the property rights system: it incentivizes investment by creating expected returns (compensation income $qd\beta\alpha$), but this is effective only when these returns exceed competitive harms ($\delta\alpha$). This conditional incentive explains why certain types of data (such as highly competitively sensitive trade secrets) struggle to incentivize sufficient investment even with property rights protection.

Proposition 3 (Systematic Underinvestment)

Even with property rights protection, when $\psi > 0$, the market equilibrium investment x_1^* is still systematically lower than the socially optimal investment:

$$x_s = \frac{\psi}{2c} \quad (7)$$

The degree of underinvestment is:

$$\frac{x_1^*}{x_s} = \frac{(qd\beta - \delta)\alpha}{\psi} < 1 \quad (8)$$

This underinvestment is not accidental but stems from three market failures: (1) Externality gap—holders are unable to capture the value of innovation $\gamma\alpha x$, which flows to the entire society; (2) Competitive damage—holders bear the entire loss $\delta\alpha x$ but only receive partial compensation, thus bearing too much negative externality; and (3) Imperfect law enforcement, $qd < 1$ prevents holders from capturing the full private value, further weakening investment incentives.

3.3. Marginal Effects of Parameters and Policy Implications

To gain a deeper understanding of the impact of each parameter on equilibrium welfare, we analyze its marginal effects. These analyses are not only of theoretical value but also provide specific guidance for policy design.

Impact of innovation externalities:

$$\frac{\partial SW_1}{\partial \gamma} = \frac{(qd\beta - \delta)\alpha^2}{2c} > 0 \quad (9)$$

This suggests that an increase in the innovation externality γ always improves welfare (when there is investment). This finding provides a theoretical basis for supporting open science and data sharing – the innovation spillovers from increasing data use can improve social welfare even without changing other institutional parameters.

Impact of competition damage:

$$\frac{\partial SW_1}{\partial \delta} = \frac{\alpha^{2[\beta(qd-1)-\gamma]}}{2c} \quad (10)$$

When $(qd-1)\beta - \gamma < 0$ (typically), increasing δ reduces welfare. The economic implication of this condition is that when law enforcement is imperfect ($qd < 1$) and innovation externalities

are large, increased competition harms reduce social welfare by reducing investment. This explains the need to strike a balance between data protection and data circulation.

Impact of privacy costs:

$$\frac{\partial SW_1}{\partial \epsilon} = -\frac{(qd\beta - \delta)\alpha}{2c} < 0 \quad (11)$$

An increase in the privacy cost ϵ linearly reduces welfare, and this effect is achieved by reducing equilibrium investment. This finding emphasizes the importance of privacy-preserving technologies—reducing ϵ through technical means can promote the development of the data economy while protecting privacy.

Impact of law enforcement intensity:

$$\frac{\partial SW_1}{\partial \epsilon} = -\frac{(qd\beta - \delta)\alpha}{2c} < 0 \quad (12)$$

When $\psi > (qd\beta - \delta)\alpha$, indicating that moderate increases in law enforcement intensity improve welfare. However, this does not necessarily mean that stricter law enforcement is better. If the cost of law enforcement is considered, assuming that the cost of maintaining the intervention probability q is $\tau(q) = \frac{\tau_0 q^2}{2}$, then the socially optimal intervention probability q^* should satisfy the marginal benefit of law enforcement equal to the marginal cost of law enforcement. This implies that a moderate level of law enforcement intensity (e.g., $q \in [0.3, 0.6]$) may be optimal, while excessive law enforcement may reduce social welfare.

3.4. Data Types and Differentiated Policy

The aforementioned model reveals a core judgment logic: whether society should encourage data investment and use depends on whether it generates net benefits at the macro level (determined by the net social value coefficient $\psi = (\beta + \gamma - \delta)\alpha - \epsilon$) and whether it effectively incentivizes data holders at the micro level (determined by $qd\beta - \delta$). However, in real data

markets, the parameter combinations influencing these two judgment criteria are heterogeneous, leading to significant divergence in the incentive effects of property rights systems in different scenarios. Therefore, based on the structural relationship between four key parameters— β , γ , δ , and ε . We identify four representative data types: innovation-driven, business-driven, competition-sensitive, and privacy-sensitive. These divisions are not artificial, but rather four "governance segments" formed by the natural intersection of the model's endogenous variables and institutional boundaries. Each segment points to specific market failure mechanisms and differentiated policy tools.

The first type is innovation-driven data, characterized by $\gamma \gg \beta$, and relatively small δ . This means the data has strong knowledge spillovers at the societal level, but due to a lack of internalization mechanisms, holders receive only a minimal return. For example, while genome sequencing data has been a milestone in disease research, it's difficult to generate equivalent returns for data collectors through market mechanisms. The UK Biobank, which spent £62 million to build a genetic data platform for 500,000 people, has generated significant spillover effects in promoting basic medical research, but its payback mechanism is extremely limited. In such cases, the equilibrium investment level is far below the social optimum, with the degree of underinvestment being:

$$1 - \frac{x_1^*}{x^s} = 1 - \frac{qd\beta - \delta}{(\beta + \gamma - \delta) - \varepsilon/\alpha} \quad (13)$$

When $\gamma \rightarrow \infty$ and β is small, this ratio approaches 1, meaning that market investment may only be 10%-20% of the social optimum. Therefore, it is necessary to internalize externalities and fill the incentive gap through direct fiscal subsidies, public-private partnerships, or public data infrastructure.

The second category is business-driven data, characterized by $\beta \gg \gamma$ and a moderate δ . This type of data is used to optimize business operations. For example, e-commerce platforms use user behavior data to improve precision marketing and inventory management efficiency, and the resulting benefits can be converted into corporate profits. Alibaba's data-related business revenue exceeded 60 billion RMB in 2021, demonstrating the high efficiency of commercializing this type of data. In such cases, property rights systems can effectively provide investment incentives, and their efficiency can be measured by the following formula:

$$\text{Market Efficiency} = \frac{qd\beta - \delta}{\beta - \delta} \quad (14)$$

When this indicator approaches 1, it indicates that existing institutional incentives are sufficient, and no additional intervention is needed. However, overly strong exclusive rights may restrict data sharing. Therefore, it is recommended that institutional design maintain a moderate qd range (e.g., 0.4–0.6) to ensure investment incentives while avoiding information barriers and platform monopolies.

The third category is competition-sensitive data, determined by the criterion $\delta/\beta \rightarrow 1$, meaning that any excess returns will be lost if the data is leaked. This type of data often has highly tactical attributes, such as high-frequency trading models, undisclosed business strategies, or market pricing logic. Top quantitative funds like Renaissance Technologies consider their trading algorithms core assets because the value of their data immediately vanishes upon disclosure. In such scenarios, even if enforcement were strengthened to bring $qd \rightarrow 1$, the net incentive term $qd\beta - \delta$ could still approach zero, making investment incentives extremely fragile. Conventional property rights protection is insufficient for effective governance of this type of data. Institutional innovations such as tiered authorization, data trusts, and alliance governance are needed to achieve limited circulation without revealing core value.

The final category is privacy-sensitive data, whose core characteristic is that privacy externalities significantly outweigh positive benefits, resulting in a net social value of $\psi = (\beta + \gamma - \delta) \alpha - \varepsilon < 0$. Typical examples include data on children's online behavior, medical records, or religious identity. Even if such data holds some value for research or business, the potential risks are so significant that the optimal social strategy is to restrict its collection and use. In such scenarios, an independent privacy regulation system should be introduced, prioritizing the establishment of data minimization principles, mandatory anonymization, use restrictions, and revocable mechanisms, shifting the focus from "incentivizing development" to "regulating boundaries."

It's important to note that data types are dynamic, not static. For example, e-commerce transaction data, characterized by strong commercial attributes, may, after desensitization and structured aggregation, be transformed into public data for academic research. During this process, its parameter characteristics change: β decreases (individual share of benefits decreases), γ increases (research spillover increases), and ε decreases (privacy risks weaken), leading to a smooth evolution from a commercial-driven to an innovation-driven model. This "type migration" suggests that institutional tools should not be isolated but should be capable of dynamic adjustment. Subsequent chapters, in the "Fair Use Mechanism" section, will attempt to construct an institutional logic for explicit judgments about ψ , enabling flexible optimization of governance paths as data types evolve.

3.5 Robustness Tests under Incomplete Information

Having established the baseline results, we must acknowledge that real data markets rarely feature the complete information and perfect enforcement assumed above. Actual markets exhibit: 1) incomplete information, 2) uncertain enforcement, and 3) heterogeneous users. As a result, we explore these three key departures from our baseline assumptions.

First, consider enforcement uncertainty. Perfect detection probability q assumes courts never make mistakes—unrealistic when data use occurs behind corporate firewalls. The Uber-Waymo dispute illustrates this vividly: detecting Uber's alleged use of Waymo's self-driving data required whistleblowers and forensic analysis, not routine monitoring. When courts face detection errors, making both false positive and false negative mistakes, holders' investment incentives change even if average detection probability remains q . Specifically, investment falls to

$$x_1^* = \frac{(qd\beta - \delta)\alpha}{2c} \quad (15)$$

minus an uncertainty discount that reflects holders' risk aversion. This yields an important policy insight: consistent enforcement with a certain detection rate may incentivize more investment than sporadic crackdowns averaging the same rate. Predictability matters as much as intensity.

Second, user heterogeneity complicates optimal design. Our representative user B masks the reality that data serves diverse purposes with varying social value. Academic researchers typically generate high knowledge spillovers γ but capture low private benefits β , while commercial users show the reverse pattern. With heterogeneous users, any uniform pricing scheme (through qd) faces a dilemma: setting qd high enough to prevent excessive commercial extraction also deters valuable research use. If proportion λ of users are researchers, equilibrium investment becomes lower than with homogeneous users. This targeting problem motivates differentiated treatment based on use characteristics—precisely what fair use doctrines attempt to achieve.

$$x_1^* = \frac{[qd(\lambda\beta_R + (1 - \lambda)\beta_C) - \delta]\alpha}{2c} \quad (16)$$

Third, our model assumes users know data value ex ante. Yet many breakthrough data applications emerged through experimentation. TikTok's recommendation algorithm value became clear only through iterative development; Cambridge Analytica discovered Facebook data's political applications through trial and error. When value discovery requires use, the game structure changes holders invest based on expected value distributions while users experiment without knowing outcomes. This creates a compound inefficiency—users may over-experiment with low-value data or under-experiment with high-value opportunities, while holders face even greater uncertainty, further depressing investment. Though modeling this formally lies beyond our current scope, it explains why data markets develop intermediate institutions like data brokers (who specialize in value discovery), regulatory sandboxes (enabling controlled experimentation), and graduated access systems (revealing value through limited use).

These extensions reinforce rather than undermine our core message. Market failures not only persist but often intensify with real-world frictions. The systematic underinvestment identified in our baseline model likely understates actual inefficiencies. This strengthens the case for the institutional innovations analyzed in subsequent chapters—not as theoretical curiosities but as practical necessities for functioning data markets.

4. Welfare-based Assessment of Fair Use Defense

4.1 From Principle to Exception

A paradox exists in data governance practice: while the law protects data property rights, widespread unauthorized data use is permitted in practice. Academic research uses commercial data, news reports cite social media, and AI training uses online text—the legality of these activities is often unclear. Regulators appear to be making implicit value judgments, but what are the criteria?

The benchmark model in Chapter 3 reveals a fundamental tension in data governance. While pure property rights protection can incentivize investment, it cannot address the problem of systemic underinvestment. When the innovation externality γ significantly exceeds the private value β , market equilibrium investment systematically falls below the socially optimal level. The root cause of this market failure lies in the inability of data holders to internalize the knowledge spillovers generated by data use—they bear the full cost of investment while only capturing a fraction of the value created.

This theoretical finding raises a practical dilemma: how can we identify and address data uses whose social value outweighs private harm in practice? A deeper contradiction lies in the fact that the probability of regulatory intervention q in the benchmark model implies the logic of fair use defense. Regulators do not enforce all unauthorized uses; instead, they consider the nature and impact of the use. Regulators may choose not to intervene in data uses that have high innovation externalities but minimal competitive harm. This selective enforcement is effectively an implicit cost-benefit analysis. Regulators judge that certain uses, while unauthorized, have social value that outweighs the harm they cause.

However, this implicit mechanism has significant flaws. Market participants cannot accurately predict which uses will be tolerated, and this uncertainty increases transaction costs. Different regulators may have different judgment criteria, leading to inconsistent enforcement. More importantly, the lack of clear rules means that many valuable data uses may be abandoned due to concerns about legal risks. These problems require us to transform implicit value judgments into explicit institutional rules.

The core contribution of this chapter is to propose a framework for determining appropriate use based on incremental social welfare. By transforming complex trade-offs into calculable welfare comparisons, we make implicit value judgments explicit, providing a scientific basis for differentiated institutional design.

4.2 Theoretical Construction of Fair Use Defense

4.2.1 Basic assumptions

To establish clear judgment criteria, we first need to determine the baseline for analysis. Continuing with the game theory framework from Chapter 3, this chapter focuses on the use phase after the investment decision has been made. Consider the scenario where owner H has completed investment x and is now faced with the use decision of user B . The investment cost has become a sunk cost and no longer influences the current decision. Social planners need to compare the social welfare of allowing use versus prohibiting use.

This *ex post* analysis is an important complement to the *ex ante* analysis in Chapter 3. *Ex post* efficiency analysis focuses on the efficient allocation of given resources, while *ex ante* incentive analysis examines the dynamic effects of incentives. A complete policy analysis requires considering both dimensions simultaneously, ensuring the efficient use of resources in the current period while maintaining long-term investment incentives.

4.2.2 A Welfare-Based Derivation of the Fair Use Defense

After understanding the analytical perspective, we need to identify the various economic effects of data use. When data use is permitted, data use generates multiple economic effects.

User B's use of data generates total value:

$$V(x) = (\beta + \gamma)\alpha x \quad (17)$$

This includes private value $\beta\alpha x$ (the portion the user can appropriate) and innovation externalities $\gamma\alpha x$ (the portion that spills over to society). At the same time, data use also generates negative externalities:

$$L(x) = \delta\alpha x + \varepsilon x \quad (18)$$

This includes competitive harm to the holder $\delta\alpha x$ and privacy externality costs εx .

Combining these effects, we can calculate net social welfare. When data use is permitted, social welfare is:

$$SW_1 = V(x) - L(x) = [(\beta + \gamma - \delta)\alpha - \varepsilon]x = \psi x \quad (19)$$

Where $\psi = (\beta + \gamma - \delta)\alpha - \varepsilon$ is the net social value coefficient.

When data use is prohibited, neither new value is created, nor additional harm is incurred. Since the investment costs have already been paid and are sunk costs, they do not affect marginal decisions. Therefore:

$$SW_2 = 0 \quad (20)$$

The incremental social welfare is:

$$\Delta SW = SW_1 - SW_2 = \psi x \quad (21)$$

This analysis derives the core judgment criterion:

Proposition 4 (Fair Use Condition)

Given the investment level $x > 0$, use that strictly improves social welfare is allowed if and only if the net social value coefficient $\psi > 0$.

From Equation (21), when $x > 0$, $\Delta SW > 0$ if and only if $\psi > 0$.

This condition has an intuitive economic meaning: data use is socially justified only when the net value created by data use exceeds the social cost it incurs. This simple judgment criterion transforms the complex trade-offs of multiple interests into a comparison of a single indicator.

Strict property rights protection means no fair use exceptions, and all unauthorized use faces an expected compensation rate of qd . According to the analysis in Chapter 3, this will lead to serious underinvestment, especially for innovation-driven data. Social losses mainly come from unrealized innovation externalities. Overly relaxed fair use may weaken investment incentives and lead to reduced investment in data collection. Although it can promote the full utilization of existing data in the short term, it may reduce the supply of high-quality data in the long term. Fair use mechanisms should not be viewed as a simple weakening of property rights protection, but rather as a refined institutional design that differentiates different types of use and achieves more precise incentives across different dimensions. Especially when market equilibrium fails systemically, relying solely on uniform institutional strength (such as a unified qd) can lead to accidental killing or permissiveness. Therefore, the institutional objective should shift from "whether to use" to "how to identify what can be used," and differentiated classification becomes key.

4.3 Parameter identification and classification regulation

4.3.1 Observable behavioral parameters

While the theoretical criteria are clear, their practical application presents a key challenge. Without accurate identification of the four core parameters β , γ , δ , and ε , the criteria cannot be implemented. Therefore, an indirect path to parameter identification must be established

through a set of observable external behavioral variables. In other words, the true institutional value of the model depends on whether we can use limited but reliable real-world characteristics to infer complex and implicit economic indicators.

This paper proposes a vector $Z = (z_1, z_2, z_3, z_4, z_5)$ composed of five behavioral characteristics as the basis for inference and constructs a functional mapping relationship between it and the parameter vector $\theta = (\beta, \gamma, \delta, \epsilon)$: $\theta = f(Z) + \omega$, where ω represents the unavoidable estimation error. Each z_i is defined to target specific parameters, providing a structured and operationalizable link between the theoretical model and empirical classification.

z_1 (data type) is the starting point for identifying γ and ϵ . Different types of data have different technical attributes and legal sensitivities. For example, genomic data is more likely to lead to social research breakthroughs (higher γ), while user behavior data often involves personal privacy (higher ϵ). Whether the data is “innovation-driven” or “privacy-sensitive” directly shapes the assessment of externalities and risks, and thus influences the sign and magnitude of the net social value coefficient ψ . Innovation-driven data suggests a policy focus on promoting use and increasing γ ; privacy-sensitive data suggests stricter constraints to control ϵ .

z_2 (scale of use) primarily helps identify β and ϵ , acting as a multiplier of both value and risk. Large-scale commercial use implies higher monetization potential, increasing private benefits β . At the same time, larger scale of use increases the likelihood of privacy breaches or re-identification risks, thereby increasing ϵ . The net effect of z_2 on ψ depends on the relative strength of these two forces: a high β can raise ψ , but a sufficiently large ϵ can offset or even reverse the gain. For example, platform-level applications serving millions of users may combine substantial revenue streams with heightened privacy risks, requiring careful balancing.

z_3 (market substitutability) contributes to the assessment of δ and is also the most sensitive variable in identifying institutional boundaries. When a data use behavior is a strong functional substitute for the holder's core business, such as replicating its recommendation system or user profiling algorithms, the behavior is more likely to undermine the holder's market advantage, making δ large and possibly approaching β (i.e., $\delta \approx \beta$). In such cases, even high γ may not restore ψ to a positive value. Conversely, if the use serves unrelated industries, complements existing products, or offers highly differentiated functionality, δ will be low, allowing ψ to turn positive. This functional-substitution approach aligns with antitrust practice and avoids over-reliance on market structure metrics alone.

z_4 (degree of technological innovation) is a key discriminator for γ . γ is positive when data use leads to new model designs, new application scenarios, or drives industry technological advancements that create social value. z_4 thus helps distinguish incremental innovation from mere replication. High z_4 signals that the use generates knowledge spillovers likely to outweigh competitive harm and privacy costs, especially in basic research or open innovation contexts, where γ may be the only source of positive ψ . Indicators include the release of new models, open-source code, expanded functionalities, or infrastructure-level contributions.

z_5 (attributes of the user) influences both β and γ by indicating the user's incentive and capacity to monetize data versus generate public value. Non-profit organizations typically have low β due to weaker monetization channels but may achieve high γ through activities aimed at public welfare or foundational research. For-profit enterprises generally exhibit the opposite pattern—high β but limited γ —and when combined with high δ or ϵ , their uses may yield a negative ψ . z_5 thus functions as a “directional signal” in the value structure, helping to identify whether private benefits or public spillovers dominate.

The mapping from Z to θ integrates these behavioral indicators. β is determined by the profit orientation of the use and its scalability (z_2, z_5). γ depends on the data type and innovation output (z_1, z_4). δ is most accurately captured by functional substitutability (z_3), and ε is determined by data sensitivity and scale, mitigated by privacy-protection measures (z_1, z_2). Advanced safeguards such as differential privacy, homomorphic encryption, and federated learning can significantly reduce ε , shifting ψ upward.

While the identification mechanism provides a directional and operational tool, it cannot yet fully specify the exact form of $f(Z)$. Observable characteristics are not static and can be refined into more granular use-case categories through case accumulation and judicial precedent. Estimation error ω also constrains accuracy, for example courts and regulators may face technical knowledge gaps, interpretive biases, or resource limitations, leading to misclassification risks.

4.3.2 Fair use analysis based on data type and purpose

Building on the data type structure established in Chapter 3 and the parameter identification logic proposed in Chapter 4, this section further combines the intrinsic attributes of data (i.e., type) with the extrinsic purpose of the behavior (i.e., usage) to construct a two-dimensional analytical framework for identifying the boundaries of reasonable use. On the vertical axis, we continue the previous chapter's classification of four typical data types: innovation-driven, commercial-driven, competition-sensitive, and privacy-sensitive. On the horizontal axis, we introduce the purpose of usage, including research and education, public interest, commercial innovation, and direct competition. Different combinations of type and usage will induce distinct parameter structures, which in turn determine whether the release criterion of $\psi > 0$ is met.

For "research and education" purposes, data use has significant knowledge spillover effects, while the competitive impact and commercial benefits it brings are limited. For example, basic scientific research often uses data publicly, often through academic papers and open-source models. The γ/β ratio is higher than in general use scenarios, and δ is also low due to the non-profit nature of the use. This biases research and education use towards the $\psi > 0$ criterion. However, this judgment is not absolute. In practice, the phenomenon of "pseudo-research" is common. Some commercial entities incorporate data into their business models by affiliating with research projects and publishing non-public results, effectively circumventing licensing costs. Therefore, in institutional implementation, the judgment of scientific research use must be verified by four conditions: the institution is non-profit, the results must be publicly committed, a peer review mechanism exists, and the use is controlled within a reasonable range. Only in this way can the theoretical structure of "high γ , low δ " be implemented as a soft presumption of fair use.

When data is used in "public interest" scenarios, its social value is often amplified by the urgency and collective nature of the use, such as infectious disease tracking, disaster response, and policy evaluation. In these scenarios, γ tends to be extremely high, even approaching infinity, while the negative term formed by δ and ϵ may not be sufficient to offset the positive effects if properly regulated. For example, disease control departments mapping the spread of an epidemic based on mobile data can not only avoid significant loss of life and economic losses but also avoid market competition and affect the interests of the original data holders. However, institutional measures must prevent the abuse of "public interest" as a slogan for unrestricted use. Therefore, it is necessary to introduce a structured set of public interest judgment criteria, including clear objectives, the principles of necessity and proportionality, time limits and geographical scope, and independent ex-post oversight mechanisms. The

principle of "information self-determination," proposed by the German Federal Constitutional Court in the census case, provides important guidance for balancing this institutional tension: even when data is collected in the public interest, there must be a clear legal basis and boundaries for its use. In contrast, commercial innovation constitutes the most complex and challenging type. While it can potentially bring social benefits, it is also most likely to trigger competitive substitution and exclusion disputes. In this case, both β and γ are positive, but the magnitude of δ is highly dependent on the specific market structure and the degree of product differentiation. If the innovative behavior manifests as incremental expansion, creating new technological paths or use cases, such as developing new urban planning algorithms using traffic data, the increase in γ may far outweigh the competitive harm caused by δ , and in this case, ψ is expected to be positive. However, if the behavior is limited to replicating existing products at a lower cost, such as distilling the API output of a large model from a small model to provide only low-cost alternative services, δ may even exceed γ , making ψ unlikely to be positive. Therefore, institutional identification should focus on the critical judgment of "whether it is incremental innovation" rather than simply focusing on whether the entity is an emerging enterprise.

Finally, if the data use is competitive, especially when the user and holder are in the same market, serving the same user group, and using similar technological approaches, the competitive harm δ is close to or even equal to β , while γ is close to zero due to the lack of innovative transformation. Under this structure, ψ is significantly less than zero, and fair use protection should, in principle, be excluded. Data interoperability between e-commerce platforms and model duplication between recommendation algorithms are typical examples of this. The exception may be antitrust intervention in the context of market failure. Specifically, when a holder possesses significant market dominance and creates a barrier to

access to data resources, fair use mechanisms can serve as a means of institutional correction. In this case, fair use is no longer limited to releasing innovation externalities but is also endowed with the competition law function of correcting market structure.

It is important to note that, many data usage behaviours do not fall strictly into a single category but often exhibit mixed characteristics. For example, when a medical AI company develops diagnostic algorithms using hospital data, it may have commercial purposes ($\beta > 0$), the potential to generate new medical knowledge ($\gamma > 0$), and involve highly sensitive information ($\varepsilon > 0$). In such situations, institutional judgment requires the "dominant feature identification method": first, identifying the most deterministic parameter structure of the usage behaviour, and then assessing whether other variables are sufficient to overturn the judgment. This approach avoids the risk of formal averaging but substantive misjudgement, making it particularly suitable for emerging scenarios such as hybrid commercial and scientific research projects and public welfare and platform-based operations. At the same time, the interactive relationship between parameters cannot be ignored. In particular, the value of ε is affected by data processing methods. Employing techniques such as differential privacy and homomorphic encryption can reduce ε , thereby shifting ψ from negative to positive. This characteristic suggests that fair use institutions should not only passively tolerate certain behaviors but also actively incentivize safer and more innovative technological paths, thereby establishing a virtuous cycle of "institutional guidance, behavioral optimization, and social benefits.

4.4 Empirical Application: DeepSeek Distillation Case

4.4.1 Case Overview and Theoretical Positioning

To test the practical value of the theoretical framework and illustrate the mechanics of welfare test, we anchor the analysis in a concrete case study. In January 2025, OpenAI accused DeepSeek of improperly using its API outputs through model distillation. Model distillation⁷ collects input-output pairs through many API calls to train models with similar functionality but smaller scale. This case involves significant investment, clear private value, potential innovation externalities, and complex competitive relationships, providing an ideal scenario for testing the theoretical framework. From a data type perspective, this case involves commercially driven data—OpenAI's model outputs have high commercial value, but DeepSeek's use may constitute "transformative use," requiring a detailed analysis of whether it satisfies the condition of $\psi > 0$.

4.4.2 Parameter Identification and Quantitative Analysis

Parameter estimation begins with the most basic investment level. First, determine the investment level x . Based on multiple sources, the training cost of GPT-4 is estimated to be between \$63 million and \$100 million. Considering accumulated R&D investment and talent costs, a conservative estimate of \$80 million is adopted. For ease of analysis, the base value is normalized to 1, and each parameter is expressed as a proportion of the base value.

Next, estimate each parameter item one by one. Estimating the private value β requires consideration of multiple dimensions. DeepSeek saves approximately \$50-70 million in training costs, generates an estimated annual revenue of \$20-30 million from its product

⁷ Model distillation technology was first proposed by Hinton et al., aiming to transfer knowledge from large models to small models.

services, and generates approximately \$100 million in valuation growth from improved technological capabilities. A comprehensive assessment yields $\beta \approx 0.35$.

Innovation externalities are a key parameter in this assessment. Innovation externalities γ are reflected in three dimensions: technology diffusion, lower barriers to entry, and localized innovation. DeepSeek has 77,000 stars on GitHub⁸, lowering the barrier to entry for AI users, particularly by improving its Chinese language processing capabilities. Based on historical experience and expert assessments, $\gamma \approx 0.70-1.05$, representing 2-3 times the private value⁹. This high ratio is typical of commercially driven data transformation.

Analysis of competitive harm is the most complex, requiring consideration of multiple factors. Competitive harm δ requires detailed analysis. DeepSeek's pricing is only 3.65% of OpenAI's¹⁰,¹⁰ indicating that the two companies serve different markets. Direct customer churn is limited, but there may be demonstration and prospective impacts. There are also positive market expansion effects. The overall estimate is $\delta \approx 0.14-0.175$.

The privacy cost ε is relatively small in this case because API interactions involve technical queries rather than personal data, so we set $\varepsilon \approx 0$.

It is important to emphasize that these parameter estimates are subject to considerable uncertainty. In particular, the estimate of the innovation externality γ relies on predictions of future technology diffusion effects and may be subject to significant error. The estimate of the competition harm δ also faces the difficulty of counterfactual analysis. This uncertainty is an inherent limitation of the parameter estimation method.

⁸ The number of GitHub stars is an important indicator for measuring the influence of open-source projects. The number of 77,000 stars indicates that DeepSeek has had a significant impact on the developer community.

⁹ The quantification of innovation externalities refers to the research method of Hall et al. on R&D spillover effects.

¹⁰ The price difference reflects different market positioning: OpenAI serves high-end individual and corporate customers, while DeepSeek targets the cost-sensitive long-tail market.

4.4.3 Fair use judgement

Substituting the estimated parameters into the judgment formula, we can draw a conclusion.

Substituting the parameters into the net social value coefficient formula:

$$\psi = (\beta + \gamma - \delta) \cdot \alpha - \varepsilon$$

Baseline scenario (taking the middle value):

$$\psi = (0.35 + 0.875 - 0.1575) \cdot 1 - 0 = 1.0675 > 0$$

To ensure the robustness of the conclusion, a sensitivity analysis is necessary. The sensitivity analysis shows that even under the most conservative estimate:

$$\psi = (0.35 + 0.7 - 0.175) \cdot 1 = 0.875 > 0$$

The conclusion still holds. DeepSeek's model distillation satisfies the condition of $\psi > 0$ and should be considered fair use.

However, the DeepSeek–OpenAI dispute is used here only to show how the ψ -test works, not to give a universal answer. The case sits in the large-language-model sector, where spillovers are high and privacy costs are low; lessons may differ in fields like medical images or IoT data. Key numbers like training cost, competitive harm, innovation gain come from press reports and rough estimates, and small shifts could turn ψ from positive to negative. The study is static, so it leaves out model updates and long feedback loops that also shape welfare. Rules on data vary across China, the EU and the US, which means the same parameters could trigger different legal outcomes. Many inputs are known only after the fact; real-time actors have far less information. A fuller test of the ψ -framework will need more cases, wider parameter ranges and dynamic checks.

5 Policy Implications and Conclusion

5.1 Policy Implications

Smarter data governance rests on two linked layers that respond to each other. *Ex ante* regulation classifies data and sets a matching level of legal protection. *Ex post* liability reviews real disputes under a fair-use defense. The exchange between rule-setting and case-by-case judgement keeps innovation open and markets fair with predictable law.

5.1.1 *Ex ante* regulation: data classification and differentiated enforcement

From 3.5, we can see that significant differences in the three key parameters of innovation spillover γ , competition sensitivity δ , and privacy cost ϵ affect the size and marginal change direction of social net benefit ψ . A one-size-fits-all approach will weaken both investment incentives and reasonable use, resulting in social welfare not being at the optimal level. To this end, the law enforcement strategy should follow the data classification, using the three axes of data innovation spillover γ/β , competition sensitivity δ/β , and privacy cost ϵ as the measurement. This gives four types of data: (1) innovation-driven; (2) business-driven; (3) competition-sensitive; (4) privacy-sensitive.

The social value of innovation-driven data mainly comes from the knowledge spillover effect, and strengthening protection may inhibit the overall innovation. Therefore, the system design needs to include such data in the weak protection and compensate for the early cleaning and labeling investment by providing alternative incentives such as financial subsidies, tax deductions, and industry co-construction platforms. The "first use—later compensation" scheme can be implemented, giving priority to compensation mechanisms such as profit sharing or attribution obligations, and only impose high compensation or bans on malicious grabbers, thereby reducing the risk of innovation interruption in the process.

The positive and negative effects of commercially driven data are balanced, and it is necessary to guide holders and users to jointly internalize external value through neutral law enforcement and market-based mechanisms to form a cooperative ecosystem. Market-based authorization mechanisms such as data exchanges, standardized licensing contracts, and API markets should be encouraged to enable holders and users to internalize some of the spillover benefits in voluntary negotiations, and to disclose data pricing guidelines to gradually and reasonably convert value-added into data rents. In the law enforcement process, if the market concentration increases or the δ/β ratio rises, law enforcement agencies can curb exclusive abuse; conversely, when competition slows down or the demand for openness increases, the degree of judicial strictness can be appropriately lowered to expand data circulation.

Once competition-sensitive data is copied on a large scale without permission, it is very easy to weaken the original investors' investment intention and market advantage and trigger homogeneous competition. Therefore, a higher enforcement intensity should be adopted to enable the holder to quickly apply for a temporary injunction and request punitive damages when infringement is discovered, to form a substantial deterrent to malicious predatory use. However, for users who can prove that the social net welfare ψ is positive and does not weaken the original competitive advantage, limited fair use exemptions or compulsory licenses can be granted after review.

Privacy-sensitive data has extremely high privacy costs and strong negative externalities due to its high identifiability and potential infringement of personal interests. Therefore, the focus of governance should shift from property rights incentives to risk prevention and control, focusing on protecting the personal rights and interests of data subjects rather than investment incentives. At the institutional level, such data will be transferred to privacy-related protection frameworks such as the Personal Information Protection Law, requiring the triple

compliance obligations of "obtaining prior consent-minimum necessary-purpose limitation". If the user can prove that the data has been de-identified, the relevant desensitized data can be used under limited conditions.

As *an ex ante* regulatory mechanism, the above-mentioned differentiated law enforcement strategy is applicable to data scenarios with clear ownership and clear types. However, in practice, many usage behaviors are often in a "gray area" where the confirmation of rights is vague, or the public interest is strongly oriented. Relying solely on confirmation of rights and strong law enforcement will lead to institutional rigidity, and it is necessary to introduce *ex post* responsibility judgment for flexible supplementation.

5.1.2 Ex post liability: establishing fair use defense

The establishment of the fair use defense of data aims to liberate secondary uses of high social value from licensing barriers without weakening the incentives for future data collection and collation, thereby improving overall social welfare. In other words, if relaxing the exclusive right can bring positive marginal social benefits and will not cause the incentive for subsequent investment to fall below the threshold of $qd\beta > \delta$, then space for defense should be given in the holding rights framework.

Based on the incremental social welfare criterion proposed in Chapter 4, $\psi = (\beta + \gamma - \delta) \alpha - \epsilon$. In this expression, the five variables represent the private benefits accruing to the user (β), the social innovation externalities resulting from the behavior (γ), the market competition damage caused to the data holder (δ), the privacy risks posed to society (ϵ), and the degree to which the use behavior transforms the value of the original data (α). The fair use system addresses the question of under what combination of structures the relative weights of these variables constitute the basis for institutional release without disrupting existing investment

incentives or public order boundaries. Therefore, the fair use determination mechanism is not a value judgment between protection and restriction, but rather a dynamic assessment of the structure of interests across these five dimensions. When γ is significantly greater than β , δ is low, and α is sufficiently high, meaning that the positive externalities of the use behavior far outweigh the harm to the holder, even if unauthorized, it may be "socially acceptable" and should be released through institutional exemptions. At the institutional design level, this requires that the reasonable use system must have embedded variable-driven thinking. It cannot make static judgments based on the form of use or the type of subject but should introduce a structural trade-off mechanism.

However, these five parameters are often difficult to observe in real-world adjudications. To ensure the institutional feasibility of these criteria, they must rely on a set of observable behavioral identification mechanisms. As outlined in the identification model in Chapter 4, the institutional judgment of fair use can be inferred through five behavioral variables: data type (z_1), scale of use (z_2), market substitutability (z_3), degree of technological innovation (z_4), and subject attributes (z_5). These variables constitute an "institutional proxy structure" that both simplifies the actual judgment process and provides a basis for categorized governance. For example, if the data type is scientific research/infrastructure, the subject is a nonprofit research institution, the scale of use is moderate, and there is significant technological expansion, then it can be inferred that γ/β is high, δ and ε are low, and the overall ψ is positive. Conversely, if a commercial organization uses competitive data for similar purposes without significant transformation, even if it claims innovation, it will be difficult to meet the fair use clearance threshold. Furthermore, when these five variables are intersected with the purpose classification (research, education, public safety, commerce, etc.), they can form a "data type $\times$ purpose of use" matrix, helping the system construct soft

presumption or counter-presumption rules, thereby improving identification efficiency and transparency in adjudication and review. In other words, the design of fair use systems must move from "case-by-case judgment" to "structural modeling." By linking observable variables with parameter relationships, we can achieve an institutional synthesis of typified criteria and hierarchical authorization mechanisms.

Compared with the traditional copyright fair use system, the closed enumeration of Article 24 of China's Copyright Law and the four factors (purpose, nature, quantity, and market impact) adopted by 17 U.S.C. §107 of the United States are partially inherited in this framework: α corresponds to "quantity", γ/β and the purpose of use are nested in "nature", and δ extends "market impact". However, the data field needs to further introduce δ and ϵ , incorporate competition and privacy into a unified measurement, and make up for the traditional copyright mechanism's neglect of multiple externalities. While maintaining the connection with the traditional copyright fair use system, this design is more in line with the institutional logic of innovation, competition and privacy in China's data law governance.

5.2 Conclusion

This paper addresses the conflicting tensions between data holding rights and usage rights under the "three-rights separation" framework, asking under what conditions data users should be granted reasonable use without undermining investment incentives. Based on Gallini and Winter's (1985) "investment–infringement" two-stage game model, it develops a dynamic framework and introduces a social welfare function, $\psi = (\beta + \gamma - \delta) \alpha - \epsilon$, to measure the net marginal benefits of data usage.

The results show that granting data collectors clear resource ownership rights corrects the market failure of "zero investment" in the unprotected state, but only when expected

compensation $q\beta qd\backslash\beta\eta qd\beta$ exceeds competitive harm δ . Even with strong protection, equilibrium investment remains below the social optimum because innovation spillovers γ cannot be fully internalized providing the theoretical basis for a fair use mechanism.

To define the applicable boundaries of fair use, this article further constructs an identification mechanism between parameters and behavioral characteristics. It proposes a mapping relationship between observable variables z_1 (data type), z_2 (scale of use), z_3 (market substitutability), z_4 (technological innovation), and z_5 (user attributes) and core parameters $(\beta, \gamma, \delta, \varepsilon)$, thereby forming a two-dimensional identification framework based on "data type $\times$ purpose of use." This mechanism not only provides structural clues for the presumption of fair use in judicial adjudication but also offers theoretical support for system designers in constructing hierarchical authorization and use classification governance systems.

From a policy perspective, different types of data face distinct governance needs. For innovation-driven data, innovation externalities should be compensated through government subsidies, data openness mechanisms, or a "use first, then compensation" approach. For commercially driven data, a data authorization mechanism based on agreements should be established with appropriate enforcement intensity. For competition-sensitive data, a high-intensity protection system or methods such as data trusts and alliance governance should be adopted to reduce transaction costs. Privacy-sensitive data should be integrated into an independent privacy regulatory system, with risk control centered on ε . Fair use systems should serve as an "institutional release valve," achieving a dynamic balance between protecting exclusive rights and releasing social value through structured identification mechanisms.

However, this article still has several limitations. The modeling process assumes complete information, ignoring practical issues such as incomplete information and evidence

asymmetry in judicial practice. The mapping relationship between parameter estimates and behavioral characteristics also relies on accumulated experience and factual induction, which may lead to identification errors. Future research could consider introducing an incomplete information game framework to guide a more detailed analysis of the q-value formation mechanism. Furthermore, the proposed identification path could be empirically tested and dynamically optimized by combining AI-assisted trial mechanisms with analysis of actual judicial documents. Only by iterating between theory and practice can we optimize the data property rights governance system.

6 References

- [1] Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. "Too Much Data: Prices and Inefficiencies in Data Markets." *American Economic Journal: Microeconomics* 14, no. 4 (2022): 218-256.
- [2] Acquisti, Alessandro, Curtis Taylor, and Liad Wagman. "The Economics of Privacy." *Journal of Economic Literature* 54, no. 2 (2016): 442-492.
- [3] Alibaba Group. Annual Report 2021. Form 20-F filed with the U.S. Securities and Exchange Commission (2021).
- [4] Bao, Xiaoli, and Bingwan Xiong. "Social Relationship Capital in Contact Data: Research Paradigm of Data Factor Property Rights Allocation." *China Law Review*, no. 2 (2020): 45-67.
- [5] Bergemann, Dirk, and Alessandro Bonatti. "Markets for Information: An Introduction." *Annual Review of Economics* 11 (2019): 85-107.
- [6] Bronwyn H. Hall, Jacques Mairesse, and Pierre Mohnen, "Measuring the Returns to R&D," in *Handbook of the Economics of Innovation*, vol. 2 (2010): 1033-1082.
- [7] Chen, Yongmin. "Economic Analysis of Data Rights." *Economic Research Journal*, no. 4 (2018): 158-175.
- [8] Cheng, Xiao. "On Personal Data Rights in the Big Data Era." *Social Sciences in China*, no. 3 (2018): 102-122.
- [9] Coase, Ronald H. "The Problem of Social Cost." *Journal of Law and Economics* 3 (1960): 1-44.
- [10] Cohen, Julie E. *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford: Oxford University Press, 2019.
- [11] Contreras, Jorge L. "The New Extraterritoriality: FRAND Royalties, Anti-Suit Injunctions and the Global Race to the Bottom." *Boston University Journal of Science & Technology Law* 25, no. 2 (2019): 251-301.
- [12] Dai, Xin. "The Relational Approach to Data Rights Delineation." *Peking University Law Journal* 33, no. 6 (2021): 1612-1636.
- [13] Dosis, Anastasios, and Wilfried Sand-Zantman. "The Ownership of Data." *RAND Journal of Economics* 54, no. 1 (2023): 86-110.
- [14] Drexl, Josef. "Designing Competitive Markets for Industrial Data." *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 8, no. 4 (2017): 257-292.
- [15] Farrell, Joseph. "Information and the Coase Theorem." *Journal of Economic Perspectives* 1, no. 2 (1987): 113-129.
- [16] Fu, Xinhua. "Critique of Enterprise Data Property Rights Protection Theory: From Data Property Rights to Data Use Rights." *Oriental Law*, no. 2 (2022): 54-68.
- [17] Gallini, Nancy T., and Ralph A. Winter. "Licensing in the Theory of Innovation." *The RAND Journal of Economics* 16, no. 2 (1985): 237-252.
- [18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531* (2015).
- [19] Goldfarb, Avi, and Catherine Tucker. "Digital Economics." *Journal of Economic Literature* 57, no. 1 (2019): 3-43.
- [20] Graef, Inge. *EU Competition Law, Data Protection and Online Platforms: Data as Essential Facility*. Alphen aan den Rijn: Kluwer Law International, 2016.

- [21] He, Zhiguo, Jing Huang, and Jidong Zhou. "Open Banking: Credit Market Competition When Borrowers Own the Data." *NBER Working Paper* No. 28118 (2020).
- [22] Heller, Michael A. "The Tragedy of the Anticommons: Property in the Transition from Marx to Markets." *Harvard Law Review* 111, no. 3 (1998): 621-688.
- [23] Hohfeld, Wesley Newcomb. "Some Fundamental Legal Conceptions as Applied in Judicial Reasoning." *Yale Law Journal* 23, no. 1 (1913): 16-59.
- [24] Hu, Ling. "Formation of Data Factor Property Rights: From Legal Structure to Market Structure." *Oriental Law*, no. 2 (2022): 69-82.
- [25] IDC. "Worldwide Data Infrastructure Forecast, 2023-2027." IDC Market Analysis, Doc #US49764923 (2023).
- [26] Jin, Yao. "Reflection and Transformation of Legal Paths for Data Governance." *Legal Science*, no. 2 (2020): 71-83.
- [27] Jones, Charles I., and Christopher Tonetti. "Nonrivalry and the Economics of Data." *American Economic Review* 110, no. 9 (2020): 2819-2858.
- [28] Kerber, Wolfgang. "A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis." *GRUR International* 65, no. 11 (2016): 989-998.
- [29] Lemley, Mark A., and Carl Shapiro. "Patent Holdup and Royalty Stacking." *Texas Law Review* 85 (2007): 1991-2049.
- [30] Locke, John. *Two Treatises of Government*. Translated by Ye Qifang and Qu Junong. Beijing: Commercial Press, 1964.
- [31] Long, Weiqiu. "Research on New Data Property Rights Construction and System." *Tribune of Political Science and Law* 35, no. 4 (2017): 63-77.
- [32] Mei, Xiaying. "Between Sharing and Control: Private Law Limitations and Public Order Construction of Data Protection." *Peking University Law Journal* 31, no. 4 (2019): 863-884.
- [33] Merrill, Thomas W., and Henry E. Smith. "Optimal Standardization in the Law of Property: The Numerus Clausus Principle." *Yale Law Journal* 110, no. 1 (2000): 1-70.
- [34] Microsoft Azure. "Azure Pricing Calculator." Microsoft Corporation. <https://azure.microsoft.com/pricing/calculator/> (accessed August 5, 2025).
- [35] Penner, J. E. "The 'Bundle of Rights' Picture of Property." *UCLA Law Review* 43, no. 3 (1996): 711-820.
- [36] Qian, Ziyu. "On the Construction of Data Property Rights." *The Jurist*, no. 6 (2021): 45-58.
- [37] Shen, Weixing. "On Data Usufruct." *Social Sciences in China*, no. 11 (2020): 110-131.
- [38] Shen, Weixing, and Yun Liu. "Legislative Direction of Data Rights Confirmation." *China Cyberspace*, no. 5 (2022): 23-31.
- [39] Shen, Weixing, and Zhipeng Lu. *Theory of Data Property Rights*. Beijing: Commercial Press, 2024.
- [40] Solove, Daniel J. "The Myth of the Privacy Paradox." *George Washington Law Review* 89, no. 1 (2021): 1-51.
- [41] Stepanov, Ivan. "Introducing a Property Right Over Data in the EU: The Data Producer's Right: An Evaluation." *International Review of Law, Computers & Technology* 34, no. 1 (2020): 65-86.
- [42] Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLOS Medicine* 12, no. 3 (2015): e1001779.

- [43] Tirole, Jean. "Competition and the Industrial Challenge for the Digital Age." *Annual Review of Economics* 15 (2023): 573-605.
- [44] Varian, Hal. "Artificial Intelligence, Economics, and Industrial Organization." In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 399-419. Chicago: University of Chicago Press, 2019.
- [45] Wang, Jiandong, Shiyang Yu, and Qianqian Huang. "Research on Basic Theory and Overall Design of Data Factor System." *E-Government*, no. 2 (2022): 15-28.
- [46] Wang, Lei. "'Interrogating' Data Property Rights: From the Perspective of Information and Data Layer Division." *Journal of Huazhong University of Science and Technology (Social Sciences Edition)* 33, no. 4 (2019): 89-97.
- [47] Wang, Liming. "How to Establish Data Rights." *Chinese Journal of Law* 45, no. 4 (2023): 56-73.
- [48] Wang, Liming. "On Data Rights: From the Perspective of 'Bundle of Rights.'" *Politics and Law*, no. 7 (2022): 34-48.
- [49] Waymo LLC v. Uber Technologies, Inc., 870 F.3d 1350 (Fed. Cir. 2017).
- [50] Xu, Ke. "Data Ownership: Dual Perspectives of Economics and Law." *Electronic Intellectual Property*, no. 11 (2018): 45-58.
- [51] Xu, Ke. "Data Rights: Paradigm Integration and Normative Differentiation." *Tribune of Political Science and Law* 39, no. 4 (2021): 78-92.
- [52] Xu, Ke. "From Rights Bundle to Rights Block: Reflection and Reconstruction of Three-Rights Separation of Data." *China Law Review*, no. 2 (2023): 112-128.
- [53] Zhang, Xinbao. "On Data Property Rights as New Property Rights." *Social Sciences in China*, no. 4 (2023): 139-162.
- [54] Zhou, Linbin. "Legislative Approach to Data Rights Allocation." *People's Tribune*, no. 15 (2021): 86-89.

7 Appendix

7.1 Complete List of Model Parameters

Table 1 Model Parameters

Category	Symbol	Meaning	Value range
Decision variables	x	Data investment levels	$x \geq 0$
	x_0^*	Equilibrium investment without property rights protection	-
	x_1^*	Balanced investment under property rights protection	-
	x_s	Socially optimal investment level	-
Value Parameters	α	Basic value conversion coefficient	$\alpha > 0$
	β	Private value as a proportion of underlying value	$\beta \in (0, 1)$
	γ	Innovation externality coefficient	$\gamma > 0$
	δ	Competition damage coefficient	$\delta \in [0, 1]$
	ε	Privacy cost per unit of data	$\varepsilon \geq 0$
	c	Investment cost parameters	$c > 0$
	ψ	Net social value coefficient	$\psi = (\beta + \gamma - \delta)\alpha - \varepsilon$
Institutional parameters	q	Probability of regulatory intervention	$q \in (0, 1)$
	d	Compensation multiple	$d > 0$
	qd	Expected compensation rate	-

7.2 Technical description of user B

The representative user B represents the overall effect of N potential users in the market.

Specifically:

Total value: $V(x) = \sum v_i(x) = N\bar{v}(x) = \alpha x$, where $\bar{v}(x)$ is the average single user value

Expected compensation income: $qdV(x)$ represents the total compensation the holder receives from all infringing users

User net benefit: $(1-qd)V(x)$ represents the sum of the net benefits of all users

Due to the non-competitiveness of data, each user's decision is independent of each other, so the market equilibrium can be characterized by a representative user.

7.3 Formalization of Core Assumptions

Assumption B (using incentives): When $\beta > 0$, assume that $(1-qd)\beta > 0$, that is, $qd < 1$.

Assumption C (investment feasibility): Focus on analyzing the situation where $qd\beta > \delta$.

Technical assumption (tiny usage cost): Using data requires a tiny cost $\varepsilon' > 0$.

7.4 Detailed Derivation of Proof

Complete Proof of Proposition 1

In the absence of property rights protection ($d=0$), the optimization problem of holder H is:

$$\max_x \pi_H = -cx^2 - \delta\alpha x \quad (A1)$$

For any $x > 0$:

Investment cost: $-cx^2 < 0$

Competition damage: $-\delta\alpha x \leq 0$ (strictly less than 0 when $\delta > 0$)

Total benefit: $\pi_H < 0$

Therefore, the rational holder chooses $x_0^* = 0$, resulting in $SW_0 = SW(0) = 0$.

Complete Proof of Proposition 2

Under property rights protection ($d > 0$), the optimization problem of holder H is:

$$\max_x \pi_H = qd\beta\alpha x - cx^2 - \delta\alpha x = (qd\beta - \delta)\alpha x - cx^2 \quad (A2)$$

First-order conditions (FOC):

$$\frac{\partial \pi_H}{\partial x} = (qd\beta - \delta)\alpha - 2cx = 0 \quad (A3)$$

The solution is:

$$x = \frac{(qd\beta - \delta)\alpha}{2c} \quad (A4)$$

Second-order condition (SOC):

$$\frac{\partial^2 \pi_H}{\partial x^2} = -2c < 0 \quad (A5)$$

Confirmed to be the maximum value. When $qd\beta > \delta$, $x_1^* = \frac{(qd\beta - \delta)\alpha}{2c} > 0$; when $qd\beta \leq \delta$, $\partial \pi_H / \partial x \leq 0$ holds for all $x > 0$, so $x_1^* = 0$.

Complete Proof of Proposition 3

The social planner's optimization problem is:

$$\max_x SW = \psi x - cx^2 \quad (A6)$$

where $\psi = (\beta + \gamma - \delta)\alpha - \varepsilon$.

First-order conditions:

$$\frac{\partial SW}{\partial x} = \psi - 2cx = 0 \quad (A7)$$

Solve for the socially optimal investment:

$$x_s = \frac{\psi}{2c} \quad (A8)$$

Degree of underinvestment:

$$\frac{x_1^*}{x_s} = \frac{\left[\frac{(qd\beta - \delta)\alpha}{2c} \right]}{\left[\frac{\psi}{2c} \right]} = \frac{(qd\beta - \delta)\alpha}{\psi} \quad (A9)$$

Since $qd\beta < \beta$ (because $qd < 1$) and $\gamma > 0$, we have:

$$(qd\beta - \delta)\alpha < (\beta - \delta)\alpha < (\beta + \gamma - \delta)\alpha \leq \psi$$

Therefore, $x_1^* x_s < 1$, proving the existence of systematic underinvestment.

A.5 Detailed calculation of welfare decomposition

$x_1^* = (qd\beta - \delta)\alpha / 2c$ into each welfare component:

Private Value:

$$SW_{\{private\}(x_1^*)} = \beta \alpha x_1^* = \frac{\beta(qd\beta - \delta)\alpha^2}{2c} \quad (A10)$$

Innovation Externalities:

$$SW_{\{innovation\}(x_1^*)} = \gamma \alpha x_1^* = \frac{\gamma(qd\beta - \delta)\alpha^2}{2c} \quad (A11)$$

Competition harm:

$$L_{\{competition\}(x_1^*)} = \delta \alpha x_1^* = \frac{\delta(qd\beta - \delta)\alpha^2}{2c} \quad (A12)$$

Privacy costs:

$$E_{\{privacy\}(x_1^*)} = \varepsilon x_1^* = \frac{\varepsilon(qd\beta - \delta)\alpha}{2c} \quad (A13)$$

Investment cost:

$$C_{\{investment\}(x_1^*)} = c(x_1^*)^2 = \frac{[(qd\beta - \delta)\alpha]^2}{4c} \quad (A14)$$

Total benefits:

$$SW_1 = SW_{\{private\}} + SW_{\{innovation\}} - L_{\{competition\}} - E_{\{privacy\}} - C_{\{investment\}} \quad (A15)$$

$$S = \frac{(qd\beta - \delta)\alpha}{2c} \cdot \left[(\beta + \gamma - \delta)\alpha - \varepsilon - \frac{(qd\beta - \delta)\alpha}{2} \right] \quad (A16)$$