Data Ethics Misconduct among Researchers: Formation Process and Influencing Factors

Authors: Sun Lili, Zhang Tan, Jin Yongshi, Sun Lili

Date: 2025-07-18T00:00:00+00:00

Abstract

Abstract

[Objective] To analyze the formation process and influencing factors of data ethics misconduct among researchers in data processing and analysis, aiming to reveal the patterns of ethics misconduct formation and provide theoretical support for improving research ethics norms and optimizing research management.

[Method] Conducting grounded theory analysis of mixed data from 25 semi-structured interview transcripts and web-based topic-related texts to construct a model of the formation process and influencing factors of data ethics misconduct among researchers.

[Conclusion] Data ethics cognition constitutes the foundation for researchers' ethical judgment and norm compliance. Researchers' ethical decision-making involves a dynamic tension between deontology and utility: it may exhibit deontology-dominated "norm internalization" when norms are explicit and supervision is robust, yet may also manifest teleology-driven ethics misconduct when resource competition is intense. The structural contradictions of contextual factors and the immediate pressure of situational factors generate a resonance effect, facilitating the transformation of macro-level institutional defects into micro-level misconduct behaviors. Self-regulation and external control constitute a dual defense line, though their effectiveness is contingent upon contextual support.

Full Text

Study on the Formation Process and Influencing Factors of Data Ethical Misconduct by Researchers

Sun Lili[1], **Zhang Tan**[2], **Jin Yongshi**[2]
[1]Institute of Information Management and Technology, Nanjing University of

Technology, Nanjing 210009, China
[2]School of Economics and Management, Nanjing University of Technology, Nanjing 211816, China

Abstract:
[Objective] This study analyzes the formation process and influencing factors of data ethical misconduct among researchers during data processing and analysis, aiming to reveal the underlying patterns of ethical deviation and provide theoretical support for improving research ethics guidelines and optimizing research management. [Method] Using grounded theory analysis of mixed data from 25 semi-structured interviews and online texts related to the topic, we construct a theoretical model of the formation process and influencing factors of researchers' data ethical misconduct. [Conclusions] Data ethics cognition forms the foundation for researchers' ethical judgment and normative compliance. Researchers' ethical decision-making involves a dynamic interplay between deontological and teleological considerations: under clear norms and robust oversight, deontological imperatives lead to "norm internalization," whereas in intensely competitive resource environments, teleologically driven ethical misconduct may emerge. The resonance between structural contradictions in contextual factors and immediate pressures from situational factors facilitates the transformation of macro-level institutional deficiencies into micro-level deviant behaviors. Self-regulation and external control constitute a dual defense line, though their effectiveness depends on contextual support.

Keywords: scientific data ethics; ethical misconduct; researchers; grounded theory

Classification Number: G203

Scientific data ethics encompasses the ethical issues arising from data-related activities in scientific research, including data collection, use, description, dissemination, and openness. It represents an extension of data ethics into the research domain and overlaps with research ethics. With rapid technological advancement and the rise of data-intensive research paradigms, scientific data ethics has become a significant focus of academic attention.

Under this new research paradigm, data has become a crucial resource for scientific inquiry, yet it also presents serious ethical challenges for researchers. Data misconduct has emerged as a prominent manifestation of research ethics violations, with data-related issues accounting for 31% of all retraction incidents in recent years. Such misconduct not only undermines the research ecosystem but also erodes public trust in science. Although relevant authorities have emphasized data ethics governance and oversight of academic misconduct, issuing a series of policy documents, the intersection of these two domains—scientific data ethics—still lacks actionable management measures, and incidents continue to occur frequently. Exploring the formation process of researchers' data ethical misconduct can help reveal the deep-seated causes of such violations and provide a scientific basis for developing effective prevention and corrective measures.

Regarding the formation of ethical misconduct, deontological normative theory posits that behavior arises from internal rational concepts and a sense of moral duty, emphasizing the inherent moral value of actions themselves. Teleological normative theory, conversely, holds that an action' s morality depends primarily on its consequences, seeking to guide behavior by maximizing happiness or minimizing suffering. When making ethical decisions, individuals rarely consider only duty or consequences in isolation, but rather make choices based on a trade-off between the two. The Hunt-Vitell (H-V) model integrates these perspectives into a dual-perspective ethical evaluation framework, emphasizing that moral judgment results from the joint action of deontological principles (action legitimacy) and teleological principles (consequential utility). Compared to traditional ethical decision-making models, the H-V model' s academic value lies in its effective explanation of the complex mechanisms through which rule-oriented and utility-oriented considerations interact in ethical decision-making, its emphasis on the feedback effects of organizational environment, cultural traditions, and other contextual factors on ethical cognition, and its strong explanatory power for real-world ethical dilemmas. In the domain of research data ethics, researchers face both moral and normative constraints while bearing the pressure of uncertain outcomes—a characteristic that aligns precisely with the H-V model' s dual ethical judgment dimensions. Therefore, the H-V model provides a suitable analytical framework for this study to deconstruct the formation process of researchers' data ethical misconduct.

2.1 Big Data Ethics Research

Research on data ethics issues is quite active, with most discussions focusing on big data ethics. Key themes include: (1) manifestations of data ethics problems, encompassing personal privacy breaches, data security vulnerabilities, algorithmic discrimination, and copyright disputes; (2) studies on the causes of ethical issues, which can be summarized as the diminished agency of subjects in data usage processes, the rapid iteration and application of emerging technologies (especially big data technology), and the failure of relevant norms and regulatory measures to keep pace and form effective constraints—i.e., normative absence; (3) ethical risk identification research, where scholars have attempted to propose solutions, such as Professor Ma Haiqun' s team' s efforts to construct an intelligent intelligence analysis risk identification model aimed at early warning and identification of potential ethical risks through technical means; and (4) big data ethics governance research, which seeks to establish a multi-level governance model through collaborative governance among government, enterprises, and the public, creating a composite governance system with data security as the bottom line, privacy protection as the core, and algorithmic transparency as the safeguard to resolve the value tension between efficiency and fairness.

2.2 Research Ethics Studies

Academia has accumulated substantial research on research ethics, covering: (1) manifestations of research ethics misconduct, with existing literature identifying plagiarism, data fabrication and falsification, ethics violations, peer review fraud, and paper trading as primary forms. In the AI era, more concealed forms of misconduct have emerged, such as "intelligent plagiarism" ; (2) main causes of research ethics misconduct, with scholars using questionnaires and case analyses to explore contributing factors, finding that institutional pressure, inadequate education and cognition, defective regulatory mechanisms and ethical guidelines, and technological impact combined with profit motives are primary drivers of academic misconduct; and (3) studies on researchers' research ethics awareness, which identify issues such as conceptual misalignment, uneven cognition, and inconsistency between knowledge and action, proposing specific pathways to enhance researchers' ethical risk awareness.

2.3 Scientific Data Ethics Research

Within the scientific research domain, studies on data ethics issues remain limited, with no concentrated research themes yet emerging, though the topic has attracted academic attention. Scholars have analyzed the connotation and extension of scientific data ethics and proposed basic principles. Starting from the influencing factors of researchers' ethical behavior, they have revealed the mechanisms through which ethical cognition and attitudes affect behavioral patterns. Research has also explored ethics awareness cultivation in scientific data, proposing data ethics frameworks to guide researchers' practices and examining researchers' behavioral manifestations and underlying psychological motivations when facing data ethical dilemmas.

Literature review reveals that while big data ethics and research ethics have produced rich findings, the intersection—scientific data ethics—has received insufficient attention. Research on the causes of researchers' data misconduct is submerged within broader studies of academic integrity and misconduct, lacking targeted investigation and, more importantly, in-depth analysis of the complex psychological activities underlying the formation process of misconduct. Current data ethics research primarily focuses on big data ethics, paying less attention to scientific data with higher value density. Due to the diversity of scientific data and the characteristics of data-driven research paradigms, scientific data ethics risks possess complexity and particularity. Accordingly, this study analyzes the formation mechanism of researchers' data ethical misconduct from a process perspective, providing references for subsequent effective governance of scientific data ethics.

3.1 Research Method

Grounded theory emphasizes deriving theory from observational data rather than starting from existing theoretical assumptions. Given that current liter-

ature lacks systematic research on the formation mechanism of research data ethics misconduct and lacks mature theories for direct adoption or reference, this study employs grounded theory to conduct hierarchical coding of obtained raw materials, progressively summarizing to identify core concepts and their relationships, thereby establishing a theoretical model of the formation process of researchers' data ethical misconduct.

3.2 Data Collection

This study' s data sources primarily include two categories: semi-structured interviews and online data. To ensure high reliability and validity of the interview protocol, pilot interviews were conducted before formal interviews, and relevant expressions in the interview outline were optimized based on issues identified. The final interview outline mainly included: (1) How do you understand scientific data and the ethical issues in its processing and application in your research work? Can you provide examples from specific research scenarios? (2) Through what channels do you learn about ethics norms related to scientific data? Do these norms constrain your actual research behavior? (3) How do you implement data ethics norms in your research activities? Please describe specific scenarios; (4) When facing conflicts between scientific data ethics norms and personal interests, what factors influence your decision-making? (5) How do you explain possible misconduct by yourself or your peers? (6) How does your institution supervise and penalize data ethics violations? Have you experienced paper review questioning or revision requests due to research data ethics issues? (7) Based on your experience, what measures can effectively reduce scientific data ethics misconduct? During interviews, specific question formulations were adjusted appropriately according to interviewee roles and contexts.

Interviews were conducted primarily during May-June 2024. Following the principle of "purposive sampling," this study selected subjects who could provide maximum information relevant to the research purpose. Considering that respondents' research experience and knowledge levels affect their understanding of data ethics, selection criteria included: (1) education level of master' s degree or above; (2) experience with scientific data through contact, generation, or usage, with research experience or project involvement. Based on these criteria, three professors, two associate professors, five doctoral students, and fifteen master' s students from different disciplines were selected as interviewees. A total of 25 respondents completed formal interviews (see Table 1). With interviewees' consent, recordings were transcribed, yielding 30,000 words of interview text.

Additionally, to obtain as much qualitative material as possible while avoiding social desirability bias, this study also selected relevant user-generated content from social media as supplementary data. Public sharing of research experiences and interactive commentary on platforms like Zhihu constitutes authentic reflections of personal experiences and feelings, containing rich detail and strong explanatory power for qualitative research. This study collected and organized 904 comments from Zhihu discussions over the past five years (2019-2024) on

topics such as "Can I refuse a SCI journal' s request for raw data?" and "What are some little-known academic data fabrication behaviors around you? How do you view such behavior?" After removing duplicate and insubstantial comments, 727 valid comment entries were obtained (effective rate: 80.42%), from which 300 original statements closely related to the research theme were extracted as supplementary material to the semi-structured interviews.

4.1 Open Coding

Open coding involves labeling raw interview materials, repeatedly comparing and organizing them, extracting concepts, and inductively grouping similar concepts into categories. In this study' s open coding process, raw materials from semi-structured interviews and online comments were analyzed word-by-word. Through initial conceptualization and categorization, interview materials were reorganized to smoothly achieve categorization. To ensure coding reliability, two researchers (A and B) coded independently, then compared results, resolving disagreements through discussion or expert consultation. Concepts appearing fewer than three times that could not be categorized were eliminated. After initial deduplication, 274 original statements were obtained, abstracting 118 initial concepts and 39 categories. To illustrate the open coding process, this study presents a partial open coding table (selecting one representative original statement for each initial concept) in Table 2 .

4.2 Axial Coding

The main task of axial coding is to analyze and compare the categories identified in open coding to discover organic relationships among (sub)categories, including causal, contextual, similar, process, and semantic relationships, thereby identifying main categories. Based on this approach, this study identified 13 subcategories and 9 relationship types, ultimately summarized as 4 main categories. The main categories and their specific connotations are shown in Table 3 .

4.3 Selective Coding

Selective coding further clarifies relationships among categories based on axial coding, through repeated deliberation and reasoning, to uncover path relationships among main categories according to a storyline (see Table 4).

4.4 Theoretical Saturation and Reliability Testing

Saturation in grounded theory research is marked by coding repetition. This study tested theoretical saturation using remaining interview materials, finding no new initial concepts, categories, or typical relationships, indicating theoretical saturation had been reached. Additionally, to verify coding reliability, this study used Nvivo 14' s coding comparison function to conduct reliability testing on independently coded results from researchers A and B. Kappa coefficients

were all above 0.8. For initial concepts with disagreements, group discussion and expert consultation were used to reclassify them based on original context and theoretical framework until consensus was reached. These test results indicate that this study' s coding possesses high reliability, capable of supporting subsequent theoretical model construction.

4.5 Model Construction

Based on the above coding results, this study constructed a theoretical model of the formation process and influencing factors of researchers' data ethical misconduct, shown in Figure 1 [Figure 1: see original paper]. The model uses "deontological-teleological dual evaluation"as the core of ethical decision-making, decomposing the formation process of researchers' data ethical misconduct into: (1) a deontological evaluation path (ethical judgment): researchers form "data ethics cognition" (such as privacy protection, informed consent) through education, training, and peer exchange, triggering internal judgments of action legitimacy and forming positive ethical intentions; (2) a teleological evaluation path: researchers conduct "teleological judgments" based on expected outcomes of data behaviors (such as research efficiency, academic reputation), forming negative ethical intentions that reflect teleological weighing of utility; (3) dynamic feedback from environment and experience: according to the H-V model, situational and contextual factors influence researchers' data ethics cognition and judgment, while consequences of misconduct reinforce subsequent ethical intentions through "experience feedback" from regulation and control, creating path dependence; and (4) regulatory mechanisms for ethical misconduct: researchers suppress negative intentions through self-regulation such as compliance cost and misconduct consequence assessment; external control limits the transformation of negative intentions into misconduct through situational pressures like ethics review and social supervision. This theoretical model captures the dynamic relationship between two core elements in the formation of researchers' data ethical misconduct in data-intensive environments—the "confrontation" between negative data ethics intentions and self-regulation/external control: when negative data ethics intentions, catalyzed by environmental factors, break through the "dual defense line" of self-regulation and external control, misconduct will be continuously reinforced and fall into a vicious cycle.

5.1 Formation of Scientific Data Ethics Intentions: Dual Judgment Logic Under the H-V Model

In psychological theory, intention is defined as an individual' s internal motivation and choice capacity when facing specific goals or behaviors. Intention is considered a key factor determining whether behavior occurs and can influence its persistence and intensity. According to the H-V model, the formation of researchers' data ethics intentions is not a single-dimensional moral choice but rather the result of a dynamic interplay between deontological norms and teleological utility.

(1) Deontologically-Driven Norm Internalization

According to the H-V model, researchers' internal deontological concepts and rule consciousness drive them to learn and internalize knowledge about scientific data ethics through multiple channels, forming understandings of scientific data ethics content and principles that become the basis for ethical judgment in research processes. Data analysis reveals that researchers' ethical cognition primarily originates from education and training, peer exchange, and media dissemination. In terms of content, researchers' cognition of scientific data ethics encompasses both ethical content and principles. Content cognition mainly includes data ownership, data quality, and privacy protection. Researchers recognize that individuals, organizations, or institutions hold rights over data they create, collect, or possess, such as usage, reproduction, modification, and publication rights. Most respondents indicated they cannot share their data due to confidentiality agreements or data ownership considerations. Simultaneously, researchers recognize the necessity of ensuring data accuracy and integrity to avoid compromising research results due to quality issues. The vast majority reported adopting measures like data desensitization and anonymization to protect research subjects' privacy during personal information collection, storage, processing, and use. Additionally, the study found that researchers recognize the need to follow data ethics principles such as informed consent, fairness, and minimizing harm, which provide behavioral guidelines. When researchers realize an action complies with ethical norms, they typically form positive ethical intentions that further drive ethical behavior. When ethical behavior receives positive feedback, it further strengthens scientific data ethics cognition. This feedback may come from peer recognition, social affirmation, or internal satisfaction. This positive feedback is not limited to external recognition but includes internal fulfillment. When researchers realize their behavior meets ethical standards, they feel pride and self-respect, which in turn enhances their positive data ethics intentions: "Throughout my research career, I have never altered a single experimental data point! I feel proud of myself···" (C64).

(2) Teleologically-Driven Utility Balancing

Under the teleological utility evaluation system, researchers tend to prioritize outcome utility over ethical norms when making data ethics judgments, focusing ethical judgment and decision-making logic on short-term gains rather than the intrinsic value of ethical norms.

For example, when facing pressure from the "up-or-out" promotion system, some researchers' ethical judgments may be swayed by practical purposes, showing disregard or avoidance of ethical norms. Coding revealed that some researchers admitted, "In research, I sometimes feel that data details aren' t that important, as long as I can meet the deadline and get results published" (C78)—a typical teleological judgment: selectively reporting favorable data or embellishing results (such as modifying experimental images) to meet journal review standards and succeed in the competitive academic evaluation system. This judgment often involves strategic neglect of research ethics risks, discounting long-term academic reputation risks for short-term practical benefits. Simul-

taneously, limitations of technical tools, such as review systems' inability to effectively identify modified experimental images, further reinforce teleological tendencies, making researchers' data ethics misconduct appear as a "high-benefit, low-risk" "rational choice." This utility-first logic may satisfy individual interests in the short term but damages research data authenticity and exacerbates the "bad money drives out good" effect in the academic ecosystem. Additionally, in the current big data-driven research paradigm, there exists a noteworthy situation where researchers struggle to strictly follow traditional ethics norms under specific research contexts and technical conditions. For instance, researchers find it difficult to obtain informed consent for large-scale data, such as comment data from social media platforms. Researchers may adopt anonymization and de-identification measures to achieve research goals while ensuring data harmlessness. This situation actually reflects the lag in current research ethics norms.

In research data ethics decision-making, the dynamic interplay between deontology and teleology manifests as a continuous struggle between the internal constraints of ethical norms and the external drive of outcome utility. Deontology' s core logic treats ethical norms as non-negotiable obligations, while teleology views ethical misconduct as rational choices for achieving utility purposes. This interplay is not static opposition but evolves dynamically with context. In situations with clear ethical norms and sound supervision mechanisms, researchers tend to follow ethics guidelines like data anonymization and informed consent, demonstrating "norm internalization." In contexts of intense resource competition and utilitarian evaluation standards, teleology gains the upper hand, and researchers easily rationalize ethical misconduct. The dynamic interplay between deontology and teleology explains why some researchers with high ethical cognition may still engage in scientific data ethics misconduct.

5.2 Factors Influencing Scientific Data Ethics Intentions: Synergistic Catalysis of Context and Situation

Through analysis of interviews and online texts, we found that when describing their experiences, researchers frequently mentioned contextual occasions and socio-technical environments where behaviors occurred. Therefore, this study incorporates situational and contextual influences into theoretical construction to reveal their effects on researchers' data ethics intentions.

(1) Contextual Factors

Context in existing research mostly appears as background factors. This study' s contextual factors mainly include researchers' social environment, technology application, and individual contexts (as shown in Figure 2 [Figure 2: see original paper]). Contextual factors play important roles in researchers' data ethics decision-making.

Social environmental factors encompass four categories: organizational culture and atmosphere, promotion and evaluation systems, research resource alloca-

tion, and data standards and ethical guidelines. First, as evidenced by representative views from interviewees, organizational culture and ethical atmosphere significantly influence scientific data ethics behavior. The degree of emphasis on scientific data ethics within organizations and implementation of ethical guidelines determine researchers' sensitivity and consciousness regarding ethical issues. For example, "My advisor said that even if you can' t get results, you must never fabricate data···" (C138); "If you follow a research advisor, you can calm down and stay in the lab for just one data point···but if the advisor specializes in short-term projects for money, there' s no time for independent thinking and problem discovery···" (C158). This phenomenon aligns with ethical decision-making theory regarding organizational factors'crucial role in individual decision-making. When a positive ethical environment exists within an organization, researchers are more willing to follow ethical norms and proactively adopt data-ethical behaviors. Second, promotion and evaluation systems and research resource allocation were frequently mentioned by interviewees and online users. In the current research environment, promotion and evaluation systems often emphasize output and publication quantity, pushing some researchers to breach ethical boundaries for funding, awards, and positions, even engaging in data fabrication. Simultaneously, researchers' perception of unfair resource allocation and academia' s "Matthew effect" often prevents young scholars from obtaining adequate research resources. Additionally, coding revealed that lacking data standards and ethical guidelines may leave researchers without operational bases or ethical benchmarks during data collection, processing, and analysis. The "publish or perish" evaluation pressure, scarce research resources, and absence of ethical guidelines force researchers to make difficult choices between meeting assessment requirements and following ethical standards, continuously eroding ethical baselines.

Technological application factors refer to the application context of relevant digital technologies and system equipment in scientific research. Technical contextual factors mainly include data fusion risks, data storage and security risks, limitations of data analysis tools, and ethical risks from emerging technologies like AI during scientific data collection, processing, and analysis. Data fusion risk refers to ethical risks arising from new data and information revealed through merging multi-source data supported by big data technology. For example, even after anonymization, fused data may still re-identify individuals' sensitive information through cross-comparison, causing privacy breaches. Additionally, different data sources may contain systematic biases that, if unidentified and uncorrected, may lead to biased research findings. Coding analysis revealed that researchers' inadequacies in data storage and protection may cause data loss or inaccessibility, triggering data ethics issues. Existing research shows that research data typically becomes lost two years after publication, with availability decreasing annually. Furthermore, in data analysis, subjective data manipulation and limitations of analysis methods and tools may produce results violating ethical norms. For example, black-box characteristics of data analysis tools may lack transparency, making it difficult for researchers to understand or explain

result generation processes. Emerging technologies like AI, while enhancing data processing intelligence, also trigger new ethical risks. Some researchers mentioned that insufficient AI system transparency, difficulty guaranteeing algorithmic fairness, and inadequate data privacy protection measures constitute major ethical challenges when applying AI technology.

T.D. Wilson noted that when explaining individual information behavior mechanisms, information users should be discussed within the context of their personal information world—i.e., contextualizing the individual, emphasizing personal roles and capabilities. This study' s individual context mainly includes researchers'personality traits, data literacy levels, moral identity self-identification, and cognitive biases. Personality traits influence not only individuals'data ethics cognition but also their ethical intentions and behaviors in specific contexts. Based on coding results, researchers can be mainly categorized as conservative or aggressive. Conservative researchers tend to be more cautious, typically favoring deontological ethical evaluation during ethical judgments, and actively follow ethical norms in data management and use. Aggressive researchers are usually more adventurous, willing to break traditional ethical boundaries in research exploration to pursue benefits, testing the gray areas of scientific data ethics with relatively higher risk tolerance.

Data literacy refers to the ability to collect, process, manage, evaluate, and utilize data while complying with data ethics norms. Researchers with high data literacy possess data awareness, sensitivity, and critical thinking about data, can use data analysis tools appropriately, ensure transparency and traceability in data processing, and better identify and address scientific data ethics issues. Researchers' moral identity self-identification also plays an important role in ethical intention formation. Interview and comment texts revealed that individuals who self-identify as highly morally responsible typically follow ethical standards strictly in data management and use. Additionally, cognitive bias is an important factor influencing ethical intentions. Cognitive bias refers to people making normatively and rationally deviant judgments and decisions based on uncertain, superficial information, mainly manifesting in three aspects: (1) self-related cognitive bias: first, researchers' overconfidence leads them to overestimate their data processing capabilities and underestimate their own ethical cognition deficiencies, potentially causing them to overlook potential problems in data analysis and make unethical decisions. Second, motivated blindness and motivated reasoning further exacerbate this issue. Researchers may ignore long-term ethical misconduct consequences due to profit motives, such as "the probability of fabrication being discovered may be less than 1%" (C381). Additionally, motivated reasoning leads researchers to favor conclusions supporting their own views while ignoring relevant unfavorable data, such as "first presuppose a conclusion, then think about how to obtain data that can prove this conclusion···" (C353). When researchers have low sensitivity to "slippery slope" phenomena, they may evolve from minor issues to serious ethical misconduct, such as gradually moving from minor data alterations to systematic manipulation, with this progressive deviation causing them to gradually stray from orig-

inal ethical standards. (2) other-related cognitive bias: researchers may believe peers have similar ethical problems, thereby rationalizing their own unethical behavior, such as "if fabrication is defined as selectively reporting good-looking data, then all researchers have fabricated data···" (C211). (3) world-related cognitive bias also influences researchers' ethical judgments to some extent, such as underestimating risks of data ethics misconduct and optimistically believing problems won' t be discovered, thereby neglecting the ethical nature of behavior itself, such as "if you PS an image, reviewers can hardly redo your experiment" (C349).

(2) Situational Triggers

Situational triggers refer to external or internal stimuli in specific scenarios that prompt researchers to form negative ethical intentions, including research situations and temporal situations, reflecting immediate pressures in micro-level research contexts. According to coding materials, most researchers reported frequently facing various data challenges in research practice, including large data volumes, complex data sources and structures, privacy-sensitive data, and difficulty obtaining valuable data. Difficulty in scientific data collection and processing affects scientific data ethics intentions. For example, "because the data volume is huge, there are errors during annotation, and because the workload is too large, we won' t go back to check···" (P16). Additionally, due to disciplinary differences, certain fields like medicine and social sciences have stricter data ethics requirements due to research subject sensitivity, personal privacy risks, and social/individual impacts, making them more prone to triggering ethical issues.

Furthermore, coding analysis found that research time constraints and various career pressures are important situational triggers for negative ethical intentions, prompting researchers to focus more on research outcomes than ethical compliance when facing tight deadlines, causing ethical judgments to be compressed or ignored. For example, "completing experiments and data analysis in such a short time is too rushed, leaving no time for comprehensive data checking" (B58). Some respondents indicated that when facing pressure to complete research quickly, high-pressure review and revision processes, and "lengthy" review cycles, they might adopt non-normative behaviors due to time constraints, such as making unethical operations in data processing, excessive embellishment, or adjusting data to meet journal standards (B73, B98).

Contextual factors reflect background elements of researchers' ethical decision-making, while situational factors focus on immediate pressures. Their synergistic catalysis triggers researchers' negative ethical intentions and misconduct. The "short, flat, fast" research atmosphere, "up-or-out" evaluation mechanisms, and imbalanced resource allocation create structural contradictions (context) that intensify immediate pressures like time constraints (situation). This dual resonance makes it easier to induce researchers' scientific data ethics misconduct.

5.3 Dynamic Constraints of Self-Regulation and Social Control

This study identified and confirmed two different forms of regulation and control: self-regulation and social control. Specific path relationships are shown in Figure 3 [Figure 3: see original paper]. By analyzing logical relationships among raw material statements and directional relationships among coding nodes, we further extracted and identified different psychological and ethical decision-making pathways in researchers' self-regulation and external control processes.

Self-regulation refers to researchers'active, conscious regulation and constraint of their ethical behavior. When researchers believe that complying with scientific data ethics norms entails high time and financial costs, they may tend to simplify or circumvent some ethical requirements. Risk assessment refers to researchers' identification and analysis of potential consequences from scientific data ethics misconduct, such as threats to personal academic reputation and career development. If researchers cannot clearly recognize the serious consequences of non-compliance, risk assessment bias occurs, leading to self-regulation failure and potential ethical misconduct.

External control refers to the management and supervision of scientific data ethics misconduct by external forces, mainly including organizational supervision and social supervision. Organizational supervision refers to oversight and management of ethical misconduct by research institutions, publishers, and funding agencies. For example, research institutions typically have ethics committees responsible for reviewing and approving research projects to ensure compliance with ethical norms. Through regular ethics reviews, they examine data and reports during research processes and take corrective measures when problems are discovered. Publishers and funding agencies similarly impose ethical requirements on submitted research outcomes and proposals. Social supervision refers to broad academic and public attention to and oversight of scientific data ethics. Peer review, public discussion, and media reporting can generate external pressure on researchers, prompting them to follow scientific data ethics norms.

When researchers face high ethical compliance costs, have biased risk assessments of misconduct consequences, and external supervision becomes formalistic, dual failure of internal constraints and external control occurs, unable to dissolve researchers' negative ethical intentions, thereby producing ethical misconduct. If such misconduct goes unpunished, it further strengthens researchers' cognitive biases about scientific data ethics, reinforcing a vicious cycle of misconduct and creating a "negative intention-misconduct" reinforcement loop.

6 Research Conclusions, Contributions, and Prospects

Based on the H-V ethical decision-making model and using grounded theory analysis, this study explores the formation process and influencing factors of researchers' data ethical misconduct from a process perspective. Main conclusions include: (1) Scientific data ethics cognition forms the foundation for researchers to identify ethical issues, understand ethical norms, and evaluate

them. Researchers with higher scientific data ethics cognition can typically make more reasonable ethical judgments and form positive intentions to follow ethical norms. (2) The process of scientific data ethics misconduct always involves contextualized trade-off between deontology and utility, manifesting as a state of "dynamic adjustment." In contexts with clear ethical norms and sound supervision mechanisms, researchers tend to follow ethics guidelines like data anonymization and informed consent, demonstrating "norm internalization." In contexts of intense resource competition and utilitarian evaluation standards, researchers easily rationalize ethical misconduct. During this process, individual factors such as data literacy and cognitive bias amplify or buffer contextual influences in the trade-off. (3) The structural contradictions of contextual factors and immediate pressures of situational factors constitute triggering factors for negative data ethics intentions. The "short-sightedness" of academic evaluation systems and ambiguity of scientific data ethics guidelines form a structural breeding ground for researchers' data misconduct; immediate pressures like compressed publication cycles, large-scale research data, and processing difficulties become direct triggers for misconduct. Their resonance effect is particularly critical—for example, under the "up-or-out" system (context), researchers facing paper submission deadlines (situation) are more likely to choose data manipulation for rapid publication. This finding explains how macro-level institutional defects transform into specific misconduct through micro-level situational pressures. (4) Self-regulation and external control are important defense lines against researchers' misconduct. Compliance costs and risk assessment are primary factors influencing individual regulation behaviors. Organizational supervision and social supervision are important external constraint mechanisms that create external pressure on researchers to follow data ethics norms. However, the effectiveness of self-regulation and external control highly depends on contextual support. When transparent supervision is lacking (context), researchers' self-regulation can be undermined by fluke mentality and cognitive bias; when external control becomes formalistic, misconduct is tacitly permitted through institutional indulgence.

This study' s theoretical contributions include: (1) Through interdisciplinary theoretical integration of behavioral ethics and information behavior studies, within Chinese ethical culture and academic contexts, it constructs a theoretical model of the formation process and influencing factors of researchers' data ethical misconduct, revealing formation patterns and deepening basic theoretical research on scientific data ethics. (2) By applying the classic H-V ethical decision-making model to the special context of research ethics, it analyzes the influence of situational trigger factors and contextual factors on researchers' data ethics decision-making, advancing the model from static decision-making to dynamic situational response. (3) By deconstructing the formation process of researchers' data ethical misconduct, it reveals how macro-level institutional defects in research ethics are transmitted through meso-level organizational contexts and micro-level individual behaviors to ultimately trigger misconduct, providing a theoretical roadmap for systematic governance of research ethics.

This study ensured objectivity in category extraction through dual independent coding and Cohen' s Kappa coefficient testing. However, qualitative research' s dependence on researchers' subjective judgment may still introduce potential bias. Future research could enhance coding reliability by expanding the coding team or introducing automated text analysis tools. Additionally, the degree of influence of relevant factors on researchers' data ethics decision-making warrants further exploration.

References

[1] LIANG Y, ZHENG Y P. Research on data ethics in Scientific Research Data Sharing[J]. Forum on Science and Technology in China, 2022(01): 22-27. [2] FLORIDI L, TADDEO M. What is data ethics?[J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016, 374(2083): 20160360. [3] ZHANG Y H. Academic ethics: where to start [J]. Science & Technology Review, 2022, 40(18): 95-100. [4] YI Y S. Data academic misconduct in retracted medical articles and precautionary measures[J]. Chinese Journal of Scientific and Technical Periodicals, 2020, 31(3): 276-280. [5] SHAW D, MCMASTER R, NEWHOLM T. Care and commitment in ethical consumption: an exploration of the 'attitude-behaviour gap' [J]. Journal of Business Ethics, 2016, 136(2): 251-265. [6] VERMILLION, L.J., LASSAR, W.M. & WINSOR, R.D. The Hunt-Vitell General Theory of Marketing Ethics: Can it Enhance our Understanding of Principal-Agent Relationships in Channels of Distribution? Journal of Business Ethics 41, 267-285 (2002). [7] CHEN M, LIANG Y K. A research on anonymization of personal data in open government data privacy risk control[J]. Researches in Library Science, 2021(11): 66-71. [8] MA H Q, LI J L, YU T T, et al. A framework of ethics guidelines on public data from a whole life cycle perspective[J]. Journal of Library and Information Science in Agriculture, 2023, 35(6): 29-42. [9] ZIMMER M. "But the data is already public" : on the ethics of research in facebook[J]. Ethics and information technology, 2010, 12(4): 313-325. [10] DONG J, CHENG H. Risks and control of big data technology: analysis based on the research on the ethical issues of big data in China[J]. Studies in Dialectics of Nature, 2017, 33(11): 80-85. [11] FAN J Q, XING S C. Research progress, theoretical framework and enlightenment of big data ethics[J]. Journal of Information, 2023, 42(3): 167-173. [12] ZHANG T, MA H Q. Research on the construction of data and algorithm risk identification model in intelligent intelligence analysis[J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(8): 832-844. [13] YAO L, CHEN L H, FAN D P. Value synergy mechanism for data ethics governance system[J]. Chinese Journal of Systems Science, 2024(4): 69-74. [14] FAN Y H. A systematic reflection on the tension between talent cultivation and scientific research integrity[J]. Chinese Journal of Systems Science, 2021, 29(4): 45-50. [15] MUÑOZ-CANTERO J M, ESPIÑEIRA-BELLÓN E M. Intelligent plagiarism as a misconduct in academic integrity[J]. Acta Médica Portuguesa, 2023, 37(1): 1-2. [16] WEI Y. Study on causes and combating strategies of scientific misconduct in the field of bio-medical research[J]. Chi-

nese Health Service Management, 2020, 37(4): 302-304. [17] CHEN C F, JIAO Y Q. Investigation on research ethical norms and strategies for system optimization[J]. Forum on Science and Technology in China, 2020(5): 24-31, 40. [18] LI X Z, LU X. "Constraint"and "Internalization": Research on the paths to improve the ethics consciousness of scientific researchers[J]. Studies in Science of Science, 2024, 42(3): 449-459. [19] HU L L, ZHU Y H, LI K, et al. Research on key issues of scientific data ethics[J]. China Science & Technology Resources Review, 2022, 54(01): 11-20. [20] ZHAO J, LIU Y J, ZHAO S S, et al. How scientific emotional exhaustion leads to scientific misconducts: from the perspective of ego depletion theory[J]. Science of Science and Management of S.& T., 2021, 42(2): 30-44. [21] LIU J Y, GU L P, ZHANG X Y, et al. Framework of data ethics for researchers in open research data environment[J]. Information Studies: Theory & Application, 2021, 44(2): 83-89. [22] LIANG Y. Ethical issues and their regulations of scientific research data in the Era of Big Data[J]. Library, 2023(7): 75-81. [23] LIEM G A D. Achievement and motivation[J/OL]. Educational Psychology, 2021, 41(4). [24] SHEERAN P, WEBB T L. The intention-behavior gap[J]. Social and Personality Psychology Compass, 2016, 10(9): 503-518. [25] ZHANG L, ZHANG M, ZHANG Y. Systematic review of social knowledge behaviors of Users [J]. Information Studies: Theory & Application, 2019, 42(8): 144-152. [26] VINES T H, ALBERT A Y K, ANDREW R L, et al. The availability of research data declines rapidly with article age[J]. Current biology: CB, 2014, 24(1): 94-97. [27] WILSON T D. Models in information behaviour research[J]. Journal of Documentation, 1999, 55(3): 249-270. [28] EICKHOFF C. Cognitive biases in crowdsourcing[C/OL]//Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. Marina Del Rey CA USA: ACM, 2018.

Corresponding author: Sun Lili, E-mail: llsun@njtech.edu.cn

Author Contributions Statement:
Sun Lili: Proposed the research topic, determined research ideas and design, revised and finalized the manuscript;
Zhang Tan: Conducted literature review, data collection, coding and analysis, wrote and revised the manuscript;
Jin Yongshi: Participated in data coding and manuscript proofreading.

Note: Figure translations are in progress. See original paper for figures.