INSTITUTE FOR DEFENSE ANALYSES

Biological Data Protection Decision Aid: A Proposed Framework to Identify Biological Datasets of National Security Concern

Robert Cubeta
Kristen Bishop
Ashley Farris
J. Clay Hamill
Janet Marroquin Pineda
Jay Shah

June 2024

IDA Product 3001345

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses under contract HQ0034-19-D-0001, project AI-6-5394 "Biology Data Risk Assessment Methodology" for the Director, Science & Technology Exploitation and Analytics, Maintaining Technology Advantage (MTA), Office of the Under Secretary of Defense, Research & Engineering. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The authors wish to thank, Jeff Grotte, Carly Cox, and Laura Odell for their thoughtful reviews, and the publications team, Florestine Purnell and Amberlee Mabe-Stanberry.

For More Information:
Mr. Robert L. Cubeta, Project Leader
rcubeta@ida.org, 703-575-4681
Ms. Jessica L. Stewart, Director, SFRD
jstewart@ida.org, 703-575-4530

INSTITUTE FOR DEFENSE ANALYSES

IDA Product 3001345

Biological Data Protection Decision Aid: A Proposed Framework to Identify Biological Datasets of National Security Concern

Robert Cubeta
Kristen Bishop
Ashley Farris
J. Clay Hamill
Janet Marroquin Pineda
Jay Shah

This page is intentionally left blank.

Executive Summary

The Maintaining Technology Advantage (MTA) directorate of the Office of the Undersecretary of Defense for Research & Engineering (OUSD(R&E)), Science and Technology Program Protection (STPP) guides the Department of Defense (DOD) in balancing the promotion and protection of critical and emerging technologies. MTA tasked the Institute for Defense Analyses (IDA) with: 1) developing a risk assessment methodology to assess the national security impact of strategic competitor acquisition of U.S. biological data and 2) applying the methodology to a collection of case studies.[1,2,3] Through the process of developing, executing, and socializing the risk assessment methodology, the IDA team, in coordination with MTA and stakeholders, identified a need for a method to gauge the national security relevance of a U.S. biological dataset, without requiring as in-depth a technical analysis as IDA's risk assessment methodology. In response to this need, IDA proposes the Biological Data Protection Decision Aid (BDPDA) shown in the figure that follows.

The BDPDA is a framework consisting of qualitative factors that relate to potential national security concerns and the scientific and technological value associated with sharing a biological dataset. The structuring and selection of factors for inclusion in the framework was informed by analytic insights from the case studies assessed while developing IDA's biological data risk assessment methodology, a review of federal guidance and regulations on data and technology protection (Appendix A), and discussions

---

1 Biological data is defined as "the information, including associated descriptors, derived from the structure, function, or process of a biological system(s) that is measured, collected, or aggregated for analysis." Source: The Whitehouse, "Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for Sustainable, Safe, and Secure American Bioeconomy," (Presidential Action, Washington, DC: The Whitehouse, September 12, 2022), https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/.

2 The risk assessment methodology and an unclassified example case study is documented in Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data*, IDA Paper P-33619 (Alexandria, VA: Institute for Defense Analyses, 2023).

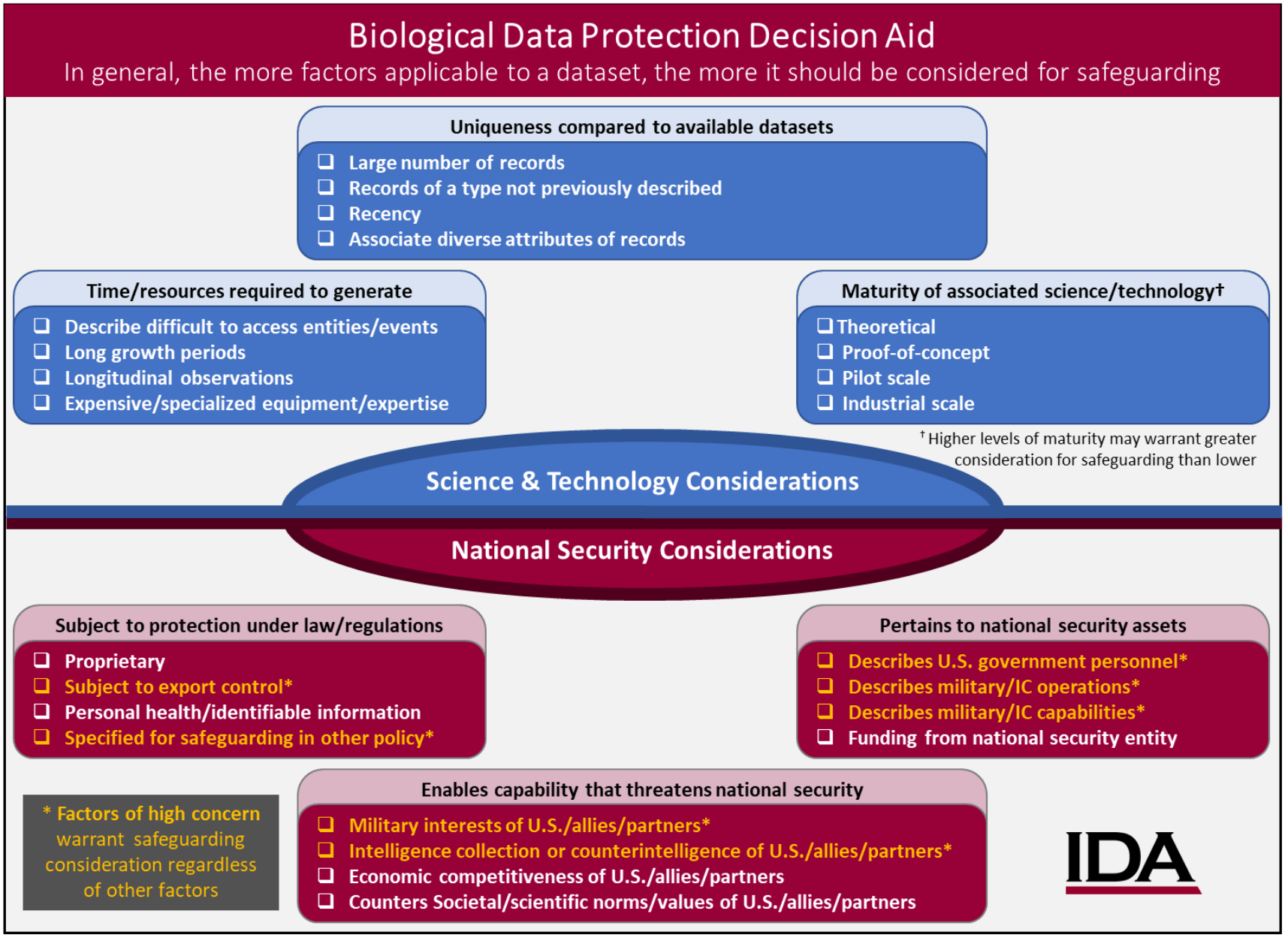
3 The complete collection of case studies can be found in Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, (U) *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*, IDA Paper P-33456 (Alexandria, VA: Institute for Defense Analyses, 2023). TOP SECRET//SI//NOFORN//FISA.

with stakeholders. The process of using the framework to identify factors that are relevant to a given dataset may facilitate:

1. Screening if the dataset warrants in-depth analysis to inform data protection decisions and
2. Communicating potential national security concerns associated with the dataset – especially between those with divergent interests (e.g., government and academia).

Use of the BDPDA on a given dataset does not generate a single definitive output or prescribe any particular actions to be taken. The framework is not intended to replace any existing processes or enforce compliance with any specific policy. Instead, the decision aid is intended to augment existing review processes and analytic tools by identifying factors that users may wish to consider during their decision making or to discuss while communicating potential national security concerns.

The intended value of the BDPDA comes through the process of reflecting on the applicability of the constituent factors. Different users may identify different sets of applicable factors. The IDA team has not yet assessed the value the BDPDA may provide to potential users. The version of the framework proposed here should be considered a work in progress until such an assessment has been conducted. An assessment of the proposed framework's value may be conducted through a workshop during which stakeholders use the BDPDA, discuss the merits of its structure and constituent factors, refine the definitions of the factors, and identify factors to remove or include. The IDA team would then update the BDPDA accordingly. Such an assessment is a critical next step in the completion of the proposed framework.



Biological Data Protection Decision Aid

This page is intentionally left blank.

Table of Contents

1. Introduction 1
2. Biological Data Protection Decision Aid 5
   A. Overview of Framework Structure 5
   B. BDPDA Constituent Factors 7
      1. Scientific and Technical Considerations 7
      2. National Security Considerations 11
3. Considerations for Implementation 15
   A. Screening if the Dataset Warrants In-depth Analysis to Inform Data Protection Decisions 15
   B. Communicating Potential National Security Concerns 16
Appendix A. Relevant Federal Policies for Biological Data Protection A-1
Appendix B. Example Application of BDPDA B-1
Appendix C. Illustrations C-1
Appendix D. References D-1
Appendix E. Abbreviations E-1

This page is intentionally left blank.

1. Introduction

The Maintaining Technology Advantage (MTA) directorate of the Office of the Undersecretary of Defense for Research & Engineering (OUSD(R&E)), Science and Technology Program Protection (STPP) guides the Department of Defense in balancing the promotion and protection of critical and emerging technologies. MTA tasked the Institute for Defense Analyses (IDA) with: 1) developing a risk assessment methodology to assess the national security impact of strategic competitor acquisition of US biological data and 2) applying the methodology to a collection of case studies.[4,5,6] Through the process of developing, executing, and socializing the risk assessment methodology, the IDA team, in coordination with MTA and stakeholders, identified a need for a method of gauging the national security relevance of a U.S. biological dataset, without requiring as in-depth a technical analysis as IDA's risk assessment methodology. In response to this need, IDA proposes the Biological Data Protection Decision Aid (BDPDA) shown in Figure 1.

The BDPDA is a framework consisting of qualitative factors relating to the potential national security concerns and the scientific and technological value of a biological dataset. The process of using the framework by identifying which factors are relevant to a given dataset may facilitate:

1. Screening if the dataset warrants in-depth analysis to inform data protection decisions and

---

4 Biological data is defined as "the information, including associated descriptors, derived from the structure, function, or process of a biological system(s) that is measured, collected, or aggregated for analysis." Source: The Whitehouse, "Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for Sustainable, Safe, and Secure American Bioeconomy," (Presidential Action, Washington, DC: The Whitehouse, September 12, 2022), https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/

5 The risk assessment methodology and an unclassified example case study are documented in Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data*, IDA Paper P-33619 (Alexandria, VA: Institute for Defense Analyses, 2023).

6 The complete collection of case studies can be found in Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, (U) *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*, IDA Paper P-33456 (Alexandria, VA: Institute for Defense Analyses, 2023). TOP SECRET//SI//NOFORN//FISA

2. Communicating potential national security concerns associated with the dataset – especially between those with divergent interests (e.g., government and academia).

Use of the BDPDA on a given dataset does not generate single definitive output or prescribe any particular actions to be taken. The framework is not intended to replace existing processes or enforce compliance with specific policy. Instead, the decision aid is intended to augment existing processes and analytic tools by identifying factors that users may wish to consider during their decision making or to highlight while communicating. The intended value of the BDPDA comes through the process of reflecting on the applicability of the constituent factors. Different users may identify different sets of applicable factors. We consider the potential for such discrepancies to be a feature, not a bug, of the framework, as they reflect the differing perspectives of the users and can serve as potential discussion points during communication or consensus building. More specifically, the BDPDA is *not* a standalone methodology to:

1. Provide a definitive characterization of risk associated with the acquisition and use of a biological dataset by a nefarious actor or strategic competitor;
2. Prescribe what, if any, specific protection measures should be taken for a particular biological dataset; or
3. Provide a risk/reward analysis between the national security risk posed by openly sharing versus the scientific or technological benefits of openly sharing a particular biological dataset.

Given that the BDPDA is not designed to generate a single definitive output for all users, its validity is not based on its ability to generate an "accurate" or reproducible result. Instead, the decision aid's validity comes from the value it provides its users. The IDA team has not yet assessed the value that the BDPDA may provide to potential users. The version of the framework proposed here should be considered a work in progress until such an assessment has been conducted. An assessment of the proposed framework's value may be conducted through a workshop during which stakeholders use the BDPDA, discuss the merits of its structure and constituent factors, refine definitions for these factors, and identify factors to remove or include. The IDA team would then update the BDPDA accordingly. Such an assessment is a critical next step in the development of the proposed framework.

The remainder of this document details the BDPDA. Chapter 2 includes discussions of the motivation for the framework's structure, the process for selecting the constituent factors, and a description of each factor. Chapter 3 contains a discussion of considerations for implementing the BDPDA. Appendix A is a brief primer on federal policies concerning biological data protection. Appendix B provides an example application of the BDPDA to a notional genetics dataset.

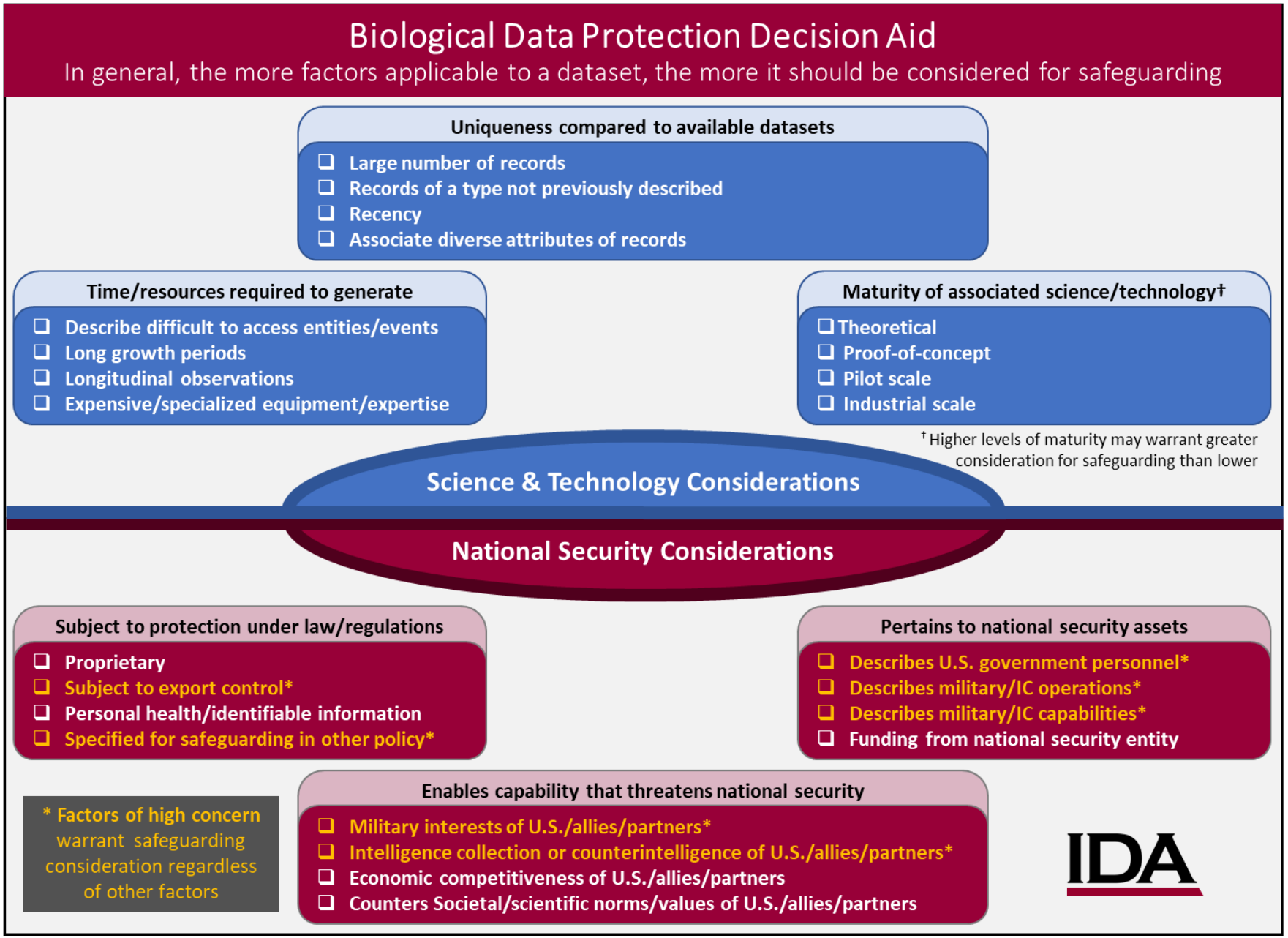


Figure 1. Biological Data Protection Decision Aid

This page is intentionally left blank.

2. Biological Data Protection Decision Aid

A. Overview of Framework Structure

As introduced in Chapter 1, the BDPDA is a structured collection of qualitative factors relating to the potential national security concerns and the scientific and technological value of a biological dataset. To execute the BDPDA, users would assess a dataset in question by considering the applicability of each factor. The structuring and selection of factors for inclusion in the framework was informed by analytic insights from the case studies assessed while developing IDA's biological data risk assessment methodology.[7]

A review of the case studies revealed that, in general, datasets with the most direct impact to near- and mid-term national security risk were those that both provide a unique value to science and technology (S&T) progression and are directly relevant to national security – be it by revealing vulnerabilities or enabling threats. As a result, the BDPDA categorizes factors as those relating to the S&T value of a dataset (top of Figure 1) or those relating to its relevance to national security (bottom of Figure 1).

The use of these two categories of factors is intended to facilitate the distinction between datasets that provide unique value to S&T advancement and innovation from those that provide S&T value **and** are directly relevant to national security. The former category, datasets that solely provide S&T value without national security implications, may not warrant data protection measures. In contrast, the latter category, datasets that provide both an S&T value and are directly relevant to national security, may warrant data protection measures. It is important to note however, that a dataset need not provide a unique S&T value (i.e., be associated with factors from the S&T consideration category) to pose a risk. In fact, certain factors in the national security considerations category may be sufficiently concerning as to warrant the protection of applicable datasets regardless of the presence of any other factors. Such factors are termed **factors of high concern**.

The factors are further grouped into three sub-categories within each of the two primary categories. The choice of the sub-categories and the specific factors contained therein was informed by the *Enablers* and *Drivers of Risk* identified in the case studies. In the risk assessment methodology, *Enablers* are capabilities and information a strategic

---

[7] Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, (U) *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*, IDA Paper P-33456 (Alexandria, VA: Institute for Defense Analyses, 2023). TOP SECRET//SI//NOFORN

competitor requires to successfully achieve a specified use of the dataset in question. *Drivers of Risk* are *Enablers* that the assessed national security risk is most sensitive to.[8] The IDA team identified *Enablers* and *Drivers of Risk* that were associated with increased risk and then generalized and grouped them into the factors and sub-categories of the BDPDA. Additional details about the sub-categories and the individual factors are included in the subsequent section.

The IDA team also reviewed federal guidance and regulations on data and technology protection. A list of relevant federal policies is included in Appendix A. The reviewed policies focus primarily on specific types, uses, and context for which biological data is regulated or protected in the United States. While these specified data types are relevant, we deemed them too specific for inclusion in a framework intended to be generally applicable to all types of biological data. Therefore, the specific types of data mentioned in these policies are intended to provide specific guidance to BDPDA users that are unknowingly generating or planning to generate biological data falling under the protected categories therein. More specifically, they are directly relevant to assessing the applicability of a dataset in question to the "subject to protection under law/regulations" factor.

In addition to drawing on existing IDA analyses, the IDA team collaborated with U.S. Government (USG) stakeholders involved in policy development for S&T protection, including those in OUSD(R&E) and the Intelligence Advanced Research Projects Activity (IARPA) to refine the list of potential factors.

The factors constituting the BDPDA are intended to be sufficiently generic as to be applicable to any type of biological data while also being sufficiently specific as to enable a meaningful assessment. Consequently, the factors are qualitative and the assessment of their applicability to a dataset requires professional judgement based on analysis and experience of the user. As previously mentioned, the framework is not intended to generate a single definitive output for all users. Instead, differing assessments of a dataset's applicability to the factors may highlight diverging view points and priorities of diverse users (e.g., government versus industry). Such variable assessments can structure discussions between the diverse users – ideally leading towards a better understanding of each other's concerns relating to the safeguarding or sharing of the dataset.

The process of using the BDPDA consists of a user evaluating a dataset in question for each factor. As the framework does not generate a prescriptive outcome, there is no prescribed mapping of a number of applicable factors to any specific data protection action.

---

[8] A complete list of each case study's *Enablers* and *Drivers of Risk* can be found in: Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, (U) *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*, IDA Paper P-33456 (Alexandria, VA: Institute for Defense Analyses, 2023). TOP SECRET//SI//NOFORN

As such, we intend for the value of the BDPDA to be process-derived not result-derived. In other words, the value of the BDPDA comes from users reflecting on the applicability of the constituent factor and subsequent discussions of which factors were or were not applicable, which may also facilitate a separate consensus building process. Ultimately, it is up to the user to determine what, if any, next steps should be taken towards protecting the dataset.

While the IDA team captured informal feedback on the factors of the framework with the sponsor and stakeholders, a formal process is warranted to finalize the BDPDA. Such feedback may be gathered during a stakeholder workshop in which participants would assess if the specific factors are relevant, able to be meaningfully assessed, and intuitively understandable with supporting documentation. Additionally, stakeholders can suggest additional factors for inclusion and ways to refine the factors' definitions. The remainder of the chapter details the specific factors that constitute the BDPDA.

B. BDPDA Constituent Factors

1. Scientific and Technical Considerations

a. Time/Resources Required to Generate

Certain biological research may require extensive time and resources to generate data, such as empirical data collection on organisms with long growth periods. For example, the agricultural bioprocessing case study the IDA team assessed with its risk assessment methodology centers on tracking disease resistance across multiple generations of livestock to ensure a genetic modification will be persevered during industrial scale-up.[9] Such information must be collected over a timescale relevant to the maturation of the agricultural species, which for the case study is on the scale of years. Another example of a time and resource extensive dataset would be one relating to an optimized industrial biomanufacturing processes, which may take three-to-ten years to fully scale-up and cost hundreds of millions of dollars.[10] Scale-up times of this duration were noted in the industrial biomanufacturing case study.[11]

---

[9] Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data*, IDA Paper P-33619 (Alexandria, VA: Institute for Defense Analyses, 2023).

[10] Jason S. Crater and Jefferson C. Lievense, "Scale-up of Industrial Microbial Processes," *FEMS Microbiology Letters* 2018, 13, fny138.

[11] Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, (U) *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*, IDA Paper P-33456 (Alexandria, VA: Institute for Defense Analyses, 2023). TOP SECRET//SI//NOFORN

Datasets requiring extensive time and resources to generate are of unique value for S&T innovation and progression because they are difficult to replicate. For this reason, a strategic competitor may be incentivized to acquire biological data that would require significant time and resources to generate to avoid the time and resource expenditure associated with independent generation of the dataset. The included factors are:

- Datasets *describing entities or events that are difficult to access* may require substantial time or resources to generate. For example, data generated from human subjects may require the time-consuming processes of obtaining volunteers and undergoing review boards. Another example is data describing an infrequently occurring biological phenomenon (e.g., a rare genotype), which may require substantial time and resources to locate opportunities to study the rare event.
- Datasets describing organisms with *long growth periods*, where the value of the data is related to observations across the lifespan or generations of an organism, such as the aforementioned data describing the inheritance of genetic modifications in livestock.
- Datasets containing *longitudinal observations*, where the value of the data is related to changes in the observations occurring over an extended period of time, such as observations of the dynamics of the human gut microbiome.
- Datasets generated with *expensive or specialized equipment or expertise* that the strategic competitor may only possess in limited quantities, such as those whose generation requires the use of a biosafety level 4 facility.

b. Uniqueness Compared to Available Datasets

Certain biological datasets may possess or be manipulated to reveal unique information that is not captured by other available data. Such uniqueness can stem from characteristics of the dataset itself, such as its size. Such datasets may provide a unique value to S&T innovations and advancement that otherwise would not be possible. The included factors are:

- Datasets containing a *large number of records* relative to other available analogous datasets may provide novel insights that lead to new capabilities.[12] The question of how large is *large* depends on the type of data and its intended use. Therefore, we do not provide specific thresholds for what constitutes *large*. Instead, users can survey publicly-available data repositories to gain an understanding of the typical range of dataset sizes for the type and use of data in

[12] For the purposes of the BDPDA, "records" refer to the entities being described by the dataset. Records are reflected by the rows in a typical flat dataset.

question.[13] Another, albeit more simplistic check, is to identify whether generating a large amount of data is a motivating factor for the creation of the dataset. The absolute size of a given dataset alone is not necessarily an indicator of its potential S&T value, but rather, the aim is to determine if the dataset sufficiently large as to enable unique insights that are otherwise not possible with smaller datasets.

- Datasets containing *records of a type not previously described* by other available datasets. Such datasets could include records of a previously unobserved type of subject, for example, information gathered on previously undiscovered microbial species. The types of datasets could also include records gathered on subjects in a unique context, for example, the behavior of a biological system in a unique environment.
- Datasets containing *recent* data may be a result of state-of-the-art data generation or collection processes or describe more temporally relevant objects or dynamics (e.g., near real-time environmental conditions). Such data may provide unique insights as compared to data from outdated processes or less-relevant time periods. The question of how recent is *recent* is context specific. The recency of data generated with state-of-the-art technologies or processes depends on the rate of technological progress for those capabilities. For example, improvements made to gene sequencing equipment over the past decade or so are substantial enough to warrant recommendations to resequencing or retire older data from reference cohorts (e.g., 1000 Genomes Project).[14] The recency of data describing more temporally relevant objects or dynamics depends on the timescale of change in the relevant phenomenon. For example, certain aspects of the human gut microbiome can change daily, whereas circulating strains of some viruses may change on a timescale of months or years.
- Datasets that *associate diverse attributes*[15] *of records* may reveal correlational relationships (e.g., gut microbiome and physiological effect) or provide novel insights from the synergy of the diverse attributes (e.g., multiomic data captured by genetic sequencing). This category of datasets includes blended data, "that is,

---

[13] An example of such a repository is the *Nucleic Acid Research's* online Molecular Biology Database collection that includes over 1,000 databases. Daniel J Rigden and Xosé M Fernández, "The 2023 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection," *Nucleic Acids Research* 51, no. D1 (January 6, 2023), https://doi.org/10.1093/nar/gkac1186.

[14] Luke Anderson-Trocmé et al., "Legacy Data Confound Genomics Studies," *Molecular Biology and Evolution* 37, no. 1 (August 30, 2019): 2–10, https://doi.org/10.1093/molbev/msz201.

[15] For the BDPDA, "attributes" refers to the variables describing each of the entities in the dataset. Attributes are reflected by the columns in a typical flat dataset.

combined sources of previously collected data [that] can improve the quality of analyses, enable new analyses, and reduce burden and cost to the public."[16] Such aggregation of data can provide unique value to S&T innovations and advancement that otherwise may not be possible through consideration of the individual data components.

c. Maturity of Associated Science and Technology

A dataset that is related to a more mature S&T application may more readily enable an operationalized capability as compared to data related to less mature S&T. While all levels of maturity are valuable to S&T development, more mature research likely has greater relevance to the advancement of specific operationalized applications that warrant safeguarding. We define the maturity levels as:[17]

- *Theoretical*: The stage at which an entity has sufficient data or scientific understanding of a phenomenon to develop a prototype algorithm or perform basic research (e.g., mathematical model of a proposed mechanism of action).
- *Proof-of-Concept*: The stage at which an entity has demonstrated the real-world feasibility of the technology – albeit in a limited context, such as in a controlled environment or within a model organism.
- *Pilot Scale*: The stage at which an entity has demonstrated limited scale-up of the technology. This includes technology that demonstrates an ability to function in a broader context, such as in multiple environments, in multiple species, or across multiple generations of a given species.
- *Industrial Scale*: The stage at which an entity is capable of deploying the technology on a large-scale, such as the commercialization and large-scale production of an organism with a desired effect.

Unlike the other categories of factors in which multiple factors could be applicable to a dataset, a dataset would likely only be applicable to one of the four development stages listed here. Therefore, a dataset associated with an *industrial scale* application may warrant more consideration for safeguarding than one associated with a *pilot scale*, *proof-of-concept*, or *theoretical* application – in that order.

---

[16] National Academies of Sciences, Engineering, and Medicine, *Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data*, (Washington, D.C.: The National Academies Press, 2024), 1, https://doi.org/10.17226/27335.

[17] The four categories of Maturity of Associated Science/Technology are adapted from IDA's updated Biological Data Risk Assessment methodology, as documented in Robert Cubeta, Kristen Bishop, Ashley Farris *et al.*, *An Update to a Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data*, IDA Product 3001344 (Alexandria, VA: Institute for Defense Analyses, 2024).

The challenges of maturing biotechnology from a theoretical concept to an industrial scale capability was a trend observed across multiple risk assessment case studies. These challenges presented themselves in various ways. For both human gut microbiome case studies, risk was constrained by the uncertainty in the scientific feasibility of maturing the relevant capabilities beyond a theoretical level. In the agricultural bioprocessing case study, short-term risk was constrained by the final step of scaling from pilot to industrial scale. In the other industrial bioprocessing case study, a driver of risk was a strategic competitor's access to information enabling pilot scale production, however, the overall risk was again constrained by the time required to scale to industrial levels and deliver the product to the end users.

2. National Security Considerations

a. Subject to Protection Under Law/Regulations

Some biological datasets may already be subject to safeguarding through existing laws or regulations, indicating an entity has already deemed the data to be sufficiently valuable to warrant protection. Such a determination may be informative for other safeguarding decisions. The specific categories of protected data included in the decision aid are:

- Datasets that are in and of themselves deemed *proprietary* or associated with a *proprietary* product, such as individual or corporate intellectual property.
- Datasets that are *subject to export control* or are associated with an export-controlled technology (see Appendix A for examples). Given that export controls are directly related to protecting national security interests, a dataset associated with this factor warrants consideration for safeguarding regardless of the presence of other factors. Therefore, this is a **factor of high concern**.
- Datasets containing *personally identifiable information (PII)* [18] (e.g., biometric data, ethnicity, or gender) or *protected health information (PHI)* [19] (e.g., medical records or physiological data)

---

[18] The U.S. Department of Labor defines PII as: "Any representation of information that permits the identity of an individual to whom information applies to be reasonably inferred by either direct or indirect means…. Information permitting the physical or online contacting of a specific individual is the same as PII." For a full description, see: U.S. Department of Labor, "Guidance on the Protection of Personal Identifiable Information," accessed November 27, 2023, https://www.dol.gov/general/ppii.

[19] Under the Standards for Privacy of Individually Identifiable Health Information i.e. the "HIPAA Privacy Rule" (45 CFR Part 160 and 45 CFR Part 164, subparts A and E), PHI is information concerning an individual that relates to "the individual's past, present, or future physical or mental health condition, the provision of health care to the individual, or the present, past, or future payment

- Datasets of a type that are specifically mentioned as requiring safeguarding in other governmental policies (see Appendix A for examples)

b. Pertains to National Security Assets

A dataset whose observations describe national security assets may reveal, or be manipulated to reveal, vulnerabilities that a strategic competitor can exploit. Often, the extent of the vulnerability depends on the timeliness of the data. Data describing the personnel and capabilities employed in ongoing military or intelligence community (IC) operations may warrant a higher consideration for safeguarding than data related to past operations or a non-operational context. As introduced in Section 2.A, the applicability of some of these factors may be sufficiently concerning as to warrant data safeguards regardless of the presence of any other factors (i.e., **factor of high concern**). Generally, these factors relate to datasets that are directly related to military or IC assets or would provide a strategic competitor a capability that threatens military or IC interests. The BDPDA includes the following:

- Datasets describing *U.S. government personnel*—be they military service members, DOD civilians/contractors, intelligence officers/agents, or politicians—may be exploited to reveal vulnerabilities that a strategic competitor could leverage to blackmail or threaten the health and safety of these individuals, which in turn may threaten mission success. This is deemed a **factor of high concern** given the direct implication of the well-being of government personnel on national security.
- Datasets describing *military/IC operations*—be they military activities (either domestic or abroad), intelligence collection activities, counterintelligence activities, or other covert/clandestine activities—may be exploited to reveal vulnerabilities that a strategic competitor could leverage to threaten mission success. This is deemed a **factor of high concern** given the direct implication of military/IC operational success on national security.
- Datasets describing *military/IC capabilities*—be they materiel or non-materiel—may be exploited to reveal vulnerabilities that a strategic competitor could leverage to threaten the performance of the capability or improve their ability to counter the capability. This is deemed a **factor of high concern** given the direct implication of military/IC capability performance on national security.

---

for the provision of health care to the individual and that identifies the individual or for which there is a reasonable basis to believe it can be used to identify the individual" which is "held or transmitted by a covered entity or its business associate, in any form or media..". For a detailed description, see: U.S. Department of Health and Human Services, "Summary of the HIPAA Privacy Rule," last updated October 19th, 2022, https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

- *Funding from a national security entity* indicates that the dataset has some relevance to national security. Users should consider traditional national security entities (e.g., DOD, IC, and Department of Homeland Security) as well as entities relevant to the economic competitiveness and wellbeing of the U.S. population (e.g., certain components within the Departments of Commerce, Energy, Agriculture, and U.S. Food and Drug Administration).

c. Enables Capability that Threatens National Security

Up to this point, all the factors used in the BDPDA have been associated with the dataset itself. This final category relates to the potential applications of a dataset. The factors specifically focus on applications of the dataset that may enable capabilities that threaten national security – that is, the dual use potential of the dataset. For our purposes, *dual use* refers to the use of a dataset in a manner that poses a potential threat to national security, often involving the use of the data in a manner unintended by those who originally generated it. Such dual use need not directly threaten U.S. national security interests to be a concern, as threats to allies and partners are also of concern.[20]

Users of the BDPDA who are not part of the national security industry may have a more difficult time assessing these factors as they may not be familiar with the possible dual use concerns associated with the data or may not have access to classified information. It is therefore essential that, to the extent permitted, national security stakeholders communicate potential dual use concerns associated with a dataset. Ideally, such discussion would occur prior to data generation to ensure safeguards are in place for the entirety of the data's life cycle. In addition, the National Security Strategy,[21] and other security policy documents (e.g., National Defense Strategy), may provide the user with foundational understanding the topics of concern to the USG and can be leveraged to inform thinking about potential threats to national security.

Development of a capability that adds risk to our national security posture may or may not directly relate to the original purpose for collecting the data, but a wide range of potential dual use purposes should be considered, to include datasets that enable the development or use of a capability that:

---

[20] The extent of concern over capabilities that threaten allies and partners depends on the strategic value of the relationship between the United States and the specific allied or partnered country. While the strategic value of any international relationship may evolve over time, certain longstanding, high-value alliances and partnerships are of particular concern. Examples include members of the North Atlantic Treaty Organization (NATO) (e.g., the United Kingdom, Germany, and the Netherlands) and those nations designated as Major Non-NATO Allies (e.g., Australia, Japan, and South Korea).

[21] The White House, *National Security Strategy*, (Washington, DC: The White House, October 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Biden-Harris-Administrations-National-Security-Strategy-10.2022.pdf.

- Creates an unintended risk to the *military interests of the U.S. or our allies and partners*. This is deemed a **factor of high concern** given the direct implication of U.S., allies, and partner military interests to national security.
- Diminishes the ability of the *U.S. and our allies and partners* to *collect intelligence*, conduct *counterintelligence*, or improves the intelligence collection efforts of a strategic competitor. This is deemed a **factor of high concern** given the direct implication of U.S.'s, allies', and partner's intelligence tactics, techniques, and procedures to national security.
- Enables the development or use of a capability that threatens the *economic competitiveness of the U.S. and our allies and partners*, such as the development of commercial products that outcompete those of the U.S., allies, and partners, resulting in a loss of market share or potential foreign dependence.
- Enables the development or use of a capability that *counters the societal and scientific norms and value of the U.S. and our allies and partners*, such as those used to violate human rights (e.g., invasion of privacy through surveillance), conduct unethical scientific research (e.g., use of non-consenting human subjects), or violate standing treaties (e.g., the Biological Weapons Convention).

3. Considerations for Implementation

As introduced in Chapter 1, use of the BDPDA on a dataset in question is designed to facilitate:

1. Screening if the dataset warrants in-depth analysis to inform data protection decisions and
2. Communicating potential national security concerns associated with the dataset – especially between those with divergent interests (e.g., government and academia).

The IDA team envisions the framework being used by a range of stakeholders, including those in government, academia, and industry. The BDPDA is intended to be applicable to a wide range of biological data types at any stage of the data lifecycle. Given the broad scope of use cases and the non-prescriptive nature of the framework, the IDA team does not propose any specific implementation concepts. Rather, we provide the following discussion on potential approaches for implementing the proposed framework. Understanding how and when stakeholders would use the framework as well as any specific processes it can be incorporated into is an important next step in the development of the BDPDA. Such information could be solicited from the same stakeholder workshop that is used to solicit feedback on the framework itself.

Ideally, data protection decisions should be made prior to the generation of the data. That way the data can proactively be protected for the duration of its lifecycle. That said, the data lifecycle is a continuous and dynamic process and data protection decisions may need to be revisited. For example, the data generation process may produce data that differs from what was expected, or existing datasets may be aggregated, disaggregated, or otherwise manipulated. Additionally, data protection decisions should be revisited in the context of new developments in relevant biotechnology and in our understanding of the capabilities and objectives of strategic competitors.

A. Screening if the Dataset Warrants In-depth Analysis to inform Data Protection Decisions

Ideally, data protection decisions should be risk-informed and supported by in-depth technical analysis, such as the outputs generated with IDA's biological data risk assessment methodology. However, such analyses can take substantial time and require both technical

and national security expertise. [22] Data protection decisions makers may not be adequately resourced to conduct such analyses on every dataset they encounter. Therefore, the BDPDA may be used to support the process of screening of datasets to flag those that may warrant in-depth analysis to inform data protection decisions. With such an approach, users would first employ the BDPDA as part of high-level assessment of the potential relevance of the dataset to national security. If through the process of assessing the applicability of BDPDA factors to a dataset in question users identify a potential national security concern, they can then proceed to use the risk assessment methodology, or other analytic methods as appropriate.

The exact process for screening if additional in-depth analysis is warranted based on the execution of the BDPDA is context-specific and will vary across groups of users. For cases in which data protection decisions are made through a formal S&T protection review processes, execution of the BDPDA could be adopted as a specific step in the existing process. On the other hand, the flexibility of the BDPDA lends itself to supporting ad-hoc decision-making processes. Regardless, it is ultimately up to the specific user to generate guidance on what decision aid outputs indicate the need for in-depth analysis. Such guidance should be informed by the user's tolerance for incorrectly ruling-out datasets that should have received additional analysis, and their tolerance for incorrectly ruling-in datasets that did not require in-depth analysis. Decision makers' tolerance for these two types of errors may be informed by their capabilities to conduct in-depth analysis. Those with greater analytic capabilities (i.e., time and expertise) may be more tolerant of erroneously flagging datasets for in-depth analysis. Whereas, those with more constrained analytic capabilities may be less tolerant.

B. Communicating Potential National Security Concerns

The relevant stakeholders for a given biological dataset may include a diverse group of individuals with diverse interests, priorities, and expectations regarding data sharing and protection. The proposed framework is intended to facilitate communicating potential national security concerns across such diverse entities. Given the qualitative nature of the factors that constitute the decision aid, different users may have different assessments of which factors are applicable to a given dataset. Such divergent outputs of the framework may reflect the variations in the perspectives of the users. Users can explain to each other their reasoning for selecting the factors they did. Such discussions can reveal common ground and areas of misalignment – ideally leading to a better understanding of the differing perspectives on the national security concerns of the dataset. In some cases, simply reaching an understanding of the differing perspectives among the stakeholders is

[22] For the case studies used to test the risk assessment methodology, it took analysts 50-150 hours to complete each analysis.

desired. However, in other cases, stakeholders may wish to come to a consensus on the national security concerns associated with a dataset.

For cases in which consensus is desired, the BDPDA may be incorporated into a structured consensus-building exercise. One such approach could be a facilitated workshop consisting of iterative rounds of participating stakeholders independently executing the BDPDA and justifying their factor selections to the rest of the participants (e.g., the Delphi method). Through the iterative discussions, participants will have the opportunity to voice their specific national security concerns and hear those of the other participants – ideally resulting in a consensus on the national security implications of the dataset.

This page is intentionally left blank.

Appendix A. Relevant Federal Policies for Biological Data Protection

Export Controls Pertaining to Biological Data and Technologies

The Bureau of Industry and Security (BIS) within the U.S. Department of Commerce oversees the Export Administration Regulations (EAR), which are the primary authority for U.S. export control regulations. The EAR includes a Commerce Control List (CCL) comprising ten broad categories of special materials and related equipment subject to export control, each of which is further divided into five product groups.[23] Category 1, Special Materials and Related Equipment, Chemicals, Microorganisms, and Toxins, is the most explicitly relevant CCL category to biological data and research. The EAR also identifies specific items of dual use concern that may require a license for export by Export Control Classification Number (ECCN). Beyond U.S. controls, biological data may be subject to multilateral controls governed by international regimes, such as the Australia Group and the Wassenaar Arrangement.[24]

Table 1 summarizes items relevant to the life sciences that the BIS considers to be of interest for export control and that may be considered biological data as defined by Executive Order 14081.[25] Therefore, the table excludes specialized equipment and includes data about biological organisms, their elements, attributes, processes, as well as whole organisms themselves. It is important to note that as biological research continues to expand into other S&T domains, export control guidance may also expand to identify additional items for protection.

---

23 "Commerce Control List," (Washington, D.C.: U.S. Department of Commerce), https://beta.bis.gov/ear, accessed 29 February 2024.

24 Kimbery Orr, *Compliance with U.S. Export Controls as a Life Science Researcher,* (U.S. Department of Commerce, no date provided), https://www.bis.doc.gov/index.php/documents/product-guidance/1107-bioexport-pdf/file, accessed 29 February 2024.

25 The Whitehouse, "Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for Sustainable, Safe, and Secure American Bioeconomy," (Presidential Action, Washington, D.C.: The Whitehouse, September 12, 2022), https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/

Table A-1. Select Biological Research Items Subject to Export Control

Control Regime	Protected Data and ECCN (as applicable) †
U.S. BIS	AG-Controlled biological agents and Select Agents under ECCN 1C351, 1C352, 1C534 †
U.S. BIS	Genetic elements for controlled agents and toxins under ECCN 1C353 †
U.S. BIS	Vaccines and medical toxins under ECCN 1C991 †
U.S. BIS	Other Biological Agents, including Ebola virus and highly pathogenic avian influenza, under 1C352a † [a]
Australia Group	Whole organisms and relevant biological data as described in AG List of Human and Animal Pathogens or Toxins [b]
Australia Group	Whole organisms and relevant biological data as described in AG List of Plant Pathogens [c]
Australia Group	Technical data for AG-controlled dual-use biological equipment such as blueprints, plans, diagrams, models, formulae, tables, engineering designs and specifications, manuals and instructions written or recorded [d]
Wassenaar Arrangement	Category 1 bioagents, biopolymers, biocatalysts [e]
Wassenaar Arrangement	Munitions List Item 7 bioagents, biopolymers, biocatalysts [e]

† Agent information can be searched by ECCN in: Bureau of Industry and Security, "CCL Index," accessed 4 January, 2024, https://beta.bis.gov/ccl-index.

[a] For additional biological agents and special material included in CCL Category 1, see: BIS, "Category 1—Special Materials and Related Equipment, Chemicals, 'Microorganisms,' and 'Toxins,'" updated 8 December 2023, https://beta.bis.gov/ear/title-15/subtitle-b/chapter-vii/subchapter-c/part-774/supplement-no-1-part-774-commerce-control#category1.

[b] The Australia Group, "List of Human and Animal Pathogens and Toxins for Export Control," updated 21 November 2023, https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/human_animal_pathogens.html.

[c] The Australia Group, "List of Plant Pathogens for Export Control," updated at 30 November 22, https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/plants.html

[d] The Australia Group, "Control List of Dual-use Biological Equipment and Related Technology and Software," updated 30 November 2022. https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/dual_biological.html

[e] Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies, *Public Documents Vol II: List of Dual-Use Goods and Technologies and Munitions List*, (December 2023), https://www.wassenaar.org/app/uploads/2023/12/List-of-Dual-Use-Goods-and-Technologies-Munitions-List-2023-1.pdf

Other Federal Policies Pertaining to Biological Data Protection

In addition to BIS-specified items subject to export control, there is a disparate collection of federal laws and regulations that relate to the protection of biological data. These policies often protect specific types of data or specific uses of data such as privacy protection for biomedical data involved in research or protection against discriminatory

practices by healthcare providers based on genetic information. Table 2 is intended to provide guidance on the types of biological data that may be considered for additional safeguarding based on their statutory status. The table is non-exhaustive but includes landmark policies that have enforcement oversight by a federal agency. Notably, these policies may specify exceptions for select types of data or uses of data. For brevity, such exceptions are not included in this table.[26]

Table A-2. Select Federal Policies Pertaining to Biological Data Protection

Federal Policy	Type of Protected Data
Health Insurance Portability and Accountability Act of 1996 (HIPAA) [a]	Patient health information collected by healthcare providers; health plans; healthcare clearinghouses; or business associates of healthcare providers, health plans (e.g., billing, claims processing)
Genetic Information Nondiscrimination Act (GINA) [b]	Genetic information; i.e., information about genetic tests, family medical history, genetic information about a fetus
Federal Policy for the Protection of Human Subjects (the "Common Rule") [c]	Research data involving human subjects, including genomic information
NIH Genomic Data Sharing Policy [d]	Individual-level, human genomic data
21st Century Cures Act (Cures Act) [e]	Identifiable biomedical information gathered or used for research purposes; biomedical data is considered identifiable when there is "at least a very small risk, as determined by current scientific practices or statistical methods" that some combination of information and other available data could be used to deduce the identity of an individual
Freedom of Information Act (FOIA) exemption	Genomic data (human)

[a] Centers for Disease Control and Prevention, "Health Insurance Portability and Accountability Act of 1996 (HIPAA)," accessed 4 January, 2024, https://www.cdc.gov/phlp/publications/topic/hipaa.html

[b] U.S. Equal Employment Opportunity Commission, "Fact Sheet: Genetic Information Nondiscrimination Act," accessed 4 January 2024, https://www.eeoc.gov/laws/guidance/fact-sheet-genetic-information-nondiscrimination-act

[c] U.S. Department of Health and Human Services, "Federal Policy for the Protection of Human Subjects ('Common Rule')," accessed 4 January 2024, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

[d] National Institutes of Health Office of Science Policy, "Genomic Data Sharing: A Two-Part Series," accessed 4 January, 2024, https://osp.od.nih.gov/genomic-data-sharing-a-two-part-series/

[e] National Human Genome Research Institute, "Laws and Regulations," accessed 4 January 2024, https://www.genome.gov/about-genomics/policy-issues/Privacy#laws-regs

[26] For more information about the scope and intent of these policies, see their corresponding references.

This page is intentionally left blank.

Appendix B. Example Application of BDPDA

This appendix explores the use of the BDPDA with an illustrative dataset and provides an example of how the BDPRA might be used to characterize national security and scientific and technological considerations. The appendix begins with a brief overview of the notional dataset before discussion of each factor that the IDA team assessed to be relevant and concludes with a short discussion.

Dataset Introduction

Privacy and national security concerns surrounding the acquisition of human genetic data have been in the forefront of recent policy decisions. In February 2024, President Biden signed the Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern, which specifically mentioned human genetic data.[27]

For these reasons, this appendix will apply the BDPDA framework to a notional human genetic dataset. The characteristics of this dataset include single nucleotide polymorphism (SNP) data obtained from 2% of the U.S. population, or approximately 6.6 million individuals. These SNP results were analyzed using the Illumina Global Screening Array Beadchip and collected from 2017 onward. In addition to the genetic data, each SNP result is associated with a profile that consists of the user's first and last name and an email address. These pieces of data are commonly collected for ancestry purposes and to find previously unknown relatives. Many direct-to-consumer (DTC) genetic testing companies collect this type of data, including 23andMe, AncestryDNA, and MyHeritage.

Relevant Factors in the BDPDA

- **Time/Resources Required to Generate**
  - **Describe difficult to access entities/events**

    A dataset of human genetic data requires soliciting volunteers and obtaining consent for the collection and storage of that data. Building a dataset

[27] "Executive Order 14117 of February 28, 2024, Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern," *Code of Federal Regulations* (2024): 15421-15430, https://www.govinfo.gov/content/pkg/FR-2024-03-01/pdf/2024-04573.pdf.

representing samples from over 6 million individuals would require extensive marketing and many years of sample collection and analysis.

- **Uniqueness Compared to Available Datasets**
  - **Large number of records**

    The dataset likely would be considered a "large" genetic dataset, with data on millions of individuals. The largest DTC genetic databases include AncestryDNA, 23andMe, and MyHeritage which have sold over 20, 12, and 5.6 million genotyping test kits, respectively.[28] The dataset examined in this appendix would be of a comparable size to the third largest DTC genetic testing dataset.

    One method of determining whether a dataset is considered large is examining new capabilities that could be developed from the dataset. One such capability is identification of individuals for genetic surveillance. While the dataset considered in this case study is not the largest DNA dataset in existence, a recent analysis by IDA estimated that a DNA dataset consisting of 2% of the U.S. population would be sufficient to provide a 10% to 40% probability of successfully identifying a target from their genetic data or that of a relative.[29] Genetic datasets with data on fewer than 100,000 individuals may be considered "small" based on the previous IDA analysis, as less than 1% of the U.S. population could be identified from a dataset of this size or smaller.

  - **Records of a type not previously described**

    This characteristic may or may not be checked for the dataset and likely will require discussion by stakeholders. One reason to not select this characteristic is because other DTC genetic testing companies possess similar data on human subjects, even if the individual customers differ. However, it also could be argued that the dataset is unique because it includes personally identifiable information from individuals not previously described.

  - **Recency**

---

[28] Ancestry, "Our History," accessed March 12, 2024, https://www.ancestry.com/corporate/about-ancestry/our-story.

23andMe, "About," accessed March 12, 2024, https://www.23andme.com/en-int/about/.

MyHeritage Blog, "MyHeritage Surpasses 1 Million Annual Subscribers," December 19, 2021, https://blog.myheritage.com/2021/12/myheritage-surpasses-1-million-annual-subscribers/.

[29] Ashley Farris and Robert Cubeta, *Security Risks Associated with the Acquisition of Human Genetic Data*, IDA Paper P-3000645 (Alexandria, VA: Institute for Defense Analyses, January 2024).

The dataset would likely be considered recent because it describes samples collected from individuals in the past 10 years. Many of the original customers would still be alive and, due to the immutability of DNA, their profiles would still be relevant.

  - **Associate diverse attributes of records**

    This dataset would likely be considered a blended dataset, as it includes both human genetic data and personally identifiable information.

- **Maturity of Associated Science and Technology**
  - **Industrial scale**

    DTC DNA testing has been conducted at a commercial industrial scale for nearly two decades, with tens of millions of test kits sold and analyzed worldwide.[30]

- **Subject to Protection Under Law/Regulations**
  - **Personal Health Information/Personally Identifiable Information**

    The dataset would be considered personally identifiable information (first name, last name, and email address), but likely will not be subject to protection as personal health information as it is currently defined. Genetic information is considered health information protected by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule following the 2013 modifications.[31] However, to be protected, it must be "individually identifiable and maintained by a covered healthcare provider, health plan, or healthcare clearinghouse."[32] Typically, genetic data collected by DTC companies is not covered under HIPAA regulation, as they are not healthcare providers, health plans, or healthcare clearinghouses.

  - **Specified for safeguarding in other policy**

    President Biden signed Executive Order 14117 in February of 2024, which states that the U.S. will restrict access to "bulk sensitive personal data",

---

[30] Rani Molla, "Why DNA tests are suddenly unpopular," Vox, February 13, 2020, https://www.vox.com/recode/2020/2/13/21129177/consumer-dna-tests-23andme-ancestry-sales-decline.

[31] "Privacy of Individually Identifiable Health Information," *Code of Federal Regulations,* title 45 subtitle A, subchapter C, part 160 (2000), https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-160, https://www.ecfr.gov/current/title-45/part-160.

[32] "Does the HIPAA Privacy Rule Protect Genetic Information?" Department of Health and Human Services, last updated December 28, 2022, https://www.hhs.gov/hipaa/for-professionals/faq/354/does-hipaa-protect-genetic-information/index.html.

which includes human genomic data, by countries of concern.[33] The Department of Justice released an Advanced Notice of Proposed Rulemaking (ANPRM) that suggested risk-based thresholds of genetic data transactions that would be restricted under the new executive order. The document suggested that human genetic data on fewer than 100 persons would comprise a low-risk transaction, whereas human genomic data on more than 1,000 persons would comprise a high-risk transaction.[34] Depending upon the final rulemaking, this additional regulation may result in restricting transactions of this nature.

- **Enables Capability that Threatens National Security**
  - **Intelligence collection or counterintelligence of U.S./Allies/Partners**

    The dataset may enable intelligence collection or counterintelligence efforts. As stated in EO 14117, malicious actors "can use their access to Americans' bulk sensitive personal data… to build profiles on United States individuals, including Federal employees and contractors, for illicit purposes, including blackmail and espionage."[35] Genetic data can reveal information that puts individuals at risk for blackmail, such as underlying health conditions or infidelity. This is a factor of high concern and would warrant additional analysis on potential risks of sharing the dataset.

  - **Economic competitiveness of U.S./Allies/Partners**

    The dataset would likely affect the economic competitiveness of the United States, allies, or partners, as genetic data is valuable in pharmaceutical and medical research. For example, in 2018 the pharmaceutical company GlaxoSmithKline invested $300 million in 23andMe to use the genetic data collected by 23andMe as the basis for the "development of innovative new

---

[33] "Executive Order 14117 of February 28, 2024, Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern," *Code of Federal Regulations* (2024): 15421-15430, https://www.govinfo.gov/content/pkg/FR-2024-03-01/pdf/2024-04573.pdf.

[34] National Security Division, Department of Justice, "Provisions Regarding Access to Americans' Bulk Sensitive Personal Data and Government-Related Data by Countries of Concern," *Code of Federal Regulations* (2024): 15780-15802, https://www.govinfo.gov/content/pkg/FR-2024-03-05/pdf/2024-04594.pdf.

[35] "Executive Order 14117 of February 28, 2024, Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern," *Code of Federal Regulations* (2024): 2, https://www.govinfo.gov/content/pkg/FR-2024-03-01/pdf/2024-04573.pdf.

medicines and potential cures." [36] Access to these diverse, large datasets could enable development of novel pharmaceuticals and medical treatments, reducing the U.S. market share in those areas.

– **Counters societal/scientific norms/values of U.S./Allies/Partners**

This dataset could be used to enable capabilities that run counter to societal and scientific norms held by the United States, particularly norms involving privacy. For years, genetic profiling of Chinese minority populations such as Uyghurs and Tibetans has been conducted by security forces as part of the passport registration process.[37] Police have used the national DNA database and other types of surveillance data to track minority groups in China.[38] With a dataset on the U.S. population, a similar capability could be developed.

- **Pertains to National Security Assets**
  - **Describes U.S. government personnel**

The dataset would likely describe U.S. government personnel, simply because of its magnitude. According to the Office of Personnel Management, there were over 1.8 million fulltime U.S. federal civilian employees in 2017.[39] Additionally, DOD reports that the active duty and reserve population in 2022 was over 2 million.[40] A dataset with the genetic data of over 6.6 million individuals is likely to contain at least some profiles of U.S. government personnel. As DNA is heritable, relatives of U.S. government personnel in the dataset could also reveal information about government personnel who never provided genetic samples themselves.

---

36 "GSK and 23andMe sign agreement to leverage genetic insights for the development of novel medicines," GSK, July 25 2018, accessed November 2023, https://www.gsk.com/en-gb/media/press-releases/gsk-and-23andme-sign-agreement-to-leverage-genetic-insights-for-the-development-of-novel-medicines/.

37 Dyani Lewis, "Unethical studies on Chinese minority groups are being retracted—but not fast enough, critics say," *Nature News* January 24, 2024, https://www.nature.com/articles/d41586-024-00170-0.

38 Yves Moreau, "Crack down on genomic surveillance," *Nature News*, December 3, 2019, https://www.nature.com/articles/d41586-019-03687-x.

39 "Federal Civilian Employment," Office of Personnel Management, September 2017, https://www.opm.gov/policy-data-oversight/data-analysis-documentation/federal-employment-reports/reports-publications/federal-civilian-employment/.

40 "Defense Department Report Shows Decline in Armed Forces Population While Percentage of Military Women Rises Slightly," U.S. Department of Defense, November 6, 2023, https://www.defense.gov/News/Releases/Release/Article/3580676/defense-department-report-shows-decline-in-armed-forces-population-while-percen/.

This is a factor of high concern and would warrant additional analysis on potential risks of sharing the dataset.

Conclusions

Overall, the dataset described in this appendix is applicable to many of the factors included in the BDPDA. Of particular note are the "intelligence collection," "describes U.S. government personnel," and "specified in other policy" factors. These factors are considered factors of high concern, which suggests that additional in-depth analysis would be needed prior to developing data protection decisions for this particular dataset.

During the process of applying the BDPDA to the dataset, we discovered that some factors may be more difficult to assess than others. For example, the dataset may be considered to have "records of a type not previously described" by some users, while others may not agree with this characterization. Applying the framework to a wide variety of dataset types with a group of diverse stakeholders and subject matter experts will undoubtedly be a valuable process as we solidify the framework and attempt to make it as useful for those generating, funding, or protecting biological datasets.

Appendix C. Illustrations

Figures

Figure 1. Biological Data Protection Decision Aid 3

Tables

Table A-1. Select Biological Research Items Subject to Export Control A-2
Table A-2. Select Federal Policies Pertaining to Biological Data Protection..... A-3

This page is intentionally left blank.

Appendix D. References

23andMe. "About." accessed March 12, 2024. https://www.23andme.com/en-int/about/.

Anderson-Trocmé, Luke, Rick Farouni, Mathieu Bourgey, Yoichiro Kamatani, Koichiro Higasa, Jeong Sun Seo, Changhoon Kim, Fumihiko Matsuda, and Simon Gravel. "Legacy Data Confound Genomics Studies." *Molecular Biology and Evolution* 37, no. 1 (August 30, 2019): 2–10. https://doi.org/10.1093/molbev/msz201.

Ancestry. "Our History." Accessed March 12, 2024. https://www.ancestry.com/corporate/about-ancestry/our-story.

The Australia Group. "List of Human and Animal Pathogens and Toxins for Export Control." Updated 21 November 2023.

The Australia Group. "Control List of Dual-use Biological Equipment and Related Technology and Software." Updated 30 November 2022.

Crater, Jason S, and Jeff Lievense. "Scale-up of Industrial Microbial Processes." *FEMS Microbiology Letters* 365, no. 13 (June 1, 2018). https://doi.org/10.1093/femsle/fny138.

Cubeta, Robert, Bishop Kristen, Marroquin Pineda, Janet, Farris, Ashley, Hammill Clay. *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data*. IDA Paper P-33619. Alexandria, VA: Institute for Defense Analyses, 2023.

Cubeta, Robert, Bishop Kristen, Marroquin Pineda, Janet, Farris, Ashley, Hammill Clay. (U) *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*. IDA Paper P-33456. Alexandria, VA: Institute for Defense Analyses, 2023. TOP SECRET//SI//NOFORN

"Defense Department Report Shows Decline in Armed Forces Population While Percentage of Military Women Rises Slightly." U.S. Department of Defense, November 6, 2023. https://www.defense.gov/News/Releases/Release/Article/3580676/defense-department-report-shows-decline-in-armed-forces-population-while-percen/.

"Does the HIPAA Privacy Rule Protect Genetic Information?" Department of Health and Human Services. Last updated December 28, 2022. https://www.hhs.gov/hipaa/for-professionals/faq/354/does-hipaa-protect-genetic-information/index.html.

Farris, Ashley, Cubeta, Robert. *Security Risks Associated with the Acquisition of Human Genetic Data*. IDA Paper P-3000645. Alexandria, VA: Institute for Defense Analyses, 2024.

"Federal Civilian Employment." Office of Personnel Management, September 2017. https://www.opm.gov/policy-data-oversight/data-analysis-documentation/federal-employment-reports/reports-publications/federal-civilian-employment/.

"GSK and 23andMe sign agreement to leverage genetic insights for the development of novel medicines." GSK, July 25 2018. Accessed November 2023. https://www.gsk.com/en-gb/media/press-releases/gsk-and-23andme-sign-agreement-to-leverage-genetic-insights-for-the-development-of-novel-medicines/.

Lewis, Dyani. "Unethical studies on Chinese minority groups are being retracted—but not fast enough, critics say." *Nature News.* January 24, 2024. https://www.nature.com/articles/d41586-024-00170-0.

Molla, Rani "Why DNA tests are suddenly unpopular." Vox, February 13, 2020. https://www.vox.com/recode/2020/2/13/21129177/consumer-dna-tests-23andme-ancestry-sales-decline.

Moreau, Yves. "Crack down on genomic surveillance." *Nature News*, December 3, 2019. https://www.nature.com/articles/d41586-019-03687-x.

National Human Genome Research Institute. "Laws and Regulations." Accessed 4 January 2024. https://www.genome.gov/about-genomics/policy-issues/Privacy#laws-regs

National Institutes of Health. *United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern* Bethesda, MD: 2012.

National Institutes of Health Office of Science Policy. "Genomic Data Sharing: A Two-Part Series." accessed 4 January, 2024. https://osp.od.nih.gov/genomic-data-sharing-a-two-part-series/

National Security Division, Department of Justice. "Provisions Regarding Access to Americans' Bulk Sensitive Personal Data and Government-Related Data by Countries of Concern." *Code of Federal Regulations,* 2024): 15780-15802, https://www.govinfo.gov/content/pkg/FR-2024-03-05/pdf/2024-04594.pdf.

Orr, Kimbery. *Compliance with U.S. Export Controls as a Life Science Researcher.* U.S. Department of Commerce: no date provided. https://www.bis.doc.gov/index.php/documents/product-guidance/1107-bioexport-pdf/file. accessed 29 February 2023.

"Privacy of Individually Identifiable Health Information." *Code of Federal Regulations,* title 45 subtitle A, subchapter C, part 160, 2000. https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-160, https://www.ecfr.gov/current/title-45/part-160.

Rigden, Daniel J., and Xosé M. Fernández. "The 2023 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection." *Nucleic Acids Research* 51, no. D1 (January 6, 2023): D1–8. https://doi.org/10.1093/nar/gkac1186.

Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data. National Academies Press eBooks, 2024. https://doi.org/10.17226/27335.

U.S. Department of Commerce. "Commerce Control List." https://beta.bis.gov/ear. accessed 29 February 2024.

U.S. Equal Employment Opportunity Commission. “Fact Sheet: Genetic Information Nondiscrimination Act.” Accessed 4 January 2024, https://www.eeoc.gov/laws/guidance/fact-sheet-genetc-information-nondiscrimination-act

U.S. Department of Health and Human Services. *Framework for Guiding Funding Decisions about Proposed Research Involving Enhanced Potential Pandemic Pathogens.* Washington, D.C.: HHS, 2017.

U.S. Department of Health and Human Services. “Federal Policy for the Protection of Human Subjects ('Common Rule').” Accessed 4 January 2024. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.

Undersecretary of Defense for Research and Engineering. *Memorandum: Policy for Risk-Based Security Reviews of Fundamental Research*. Washington, D.C.: Department of Defense, June 8, 2023.

U.S. Department of Labor. “Guidance on the Protection of Personal Identifiable Information.” Accessed November 27, 2023. https://www.dol.gov/general/ppii.

U.S. Department of Health and Human Services. “Summary of the HIPAA Privacy Rule.” Last updated October 19, 2022. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies, *Public Documents Vol II: List of Dual-Use Goods and Technologies and Munitions List.* December 2023. https://www.wassenaar.org/app/uploads/2023/12/List-of-Dual-Use-Goods-and-Technologies-Munitions-List-2023-1.pdf.

The Whitehouse. “Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for Sustainable, Safe, and Secure American Bioeconomy.” Presidential Action, Washington, D.C.: The Whitehouse, September 12, 202. https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/.

This page is intentionally left blank.

Appendix E. Abbreviations

BDPDA	Biological Data Protection Decision Aid
BIS	Bureau of Industry and Security
CFR	Code of Federal Regulations
CCL	Commerce Control List
DOD	Department of Defense
DURC	Dual Use Research of Concern
EO	Executive Order
EAR	Export Administration Regulations
ECCN	Export Control Classification Number
GINA	Genetic Information Nondiscrimination Act
HHS	Health and Human Services
HIPAA	Health Information Portability and Accountability Act
IDA	Institute for Defense Analysis
IARPA	Intelligence Advanced Research Project Activity
IC	Intelligence Community
MTA	Maintaining Technology Advantage
OUSD(R&E)	Office of the Under Secretary of Defense for Research and Engineering
PII	Personal Identifiable Information
PHI	Protected Health Information
S&T	Science and Technology
SNP	Single Nucleotide Polymorphism
STPP	Science and Technology Protection Program
US	United States

This page is intentionally left blank.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
XX-06-2024	Final	Mar 2023 - Mar 2024

4. TITLE AND SUBTITLE

Biological Data Protection Decision Aid: A Proposed Framework to Identify Biological Datasets of National Security Concern

5a. CONTRACT NO.

HQ0034-19-D-0001

5b. GRANT NO.

5c. PROGRAM ELEMENT NO(S).

6. AUTHOR(S)

Robert Cubeta
Kristen Bishop
Ashley Farris
J. Clay Hamill
Janet Marroquin-Pineda
Jay Shah

5d. PROJECT NO.

5e. TASK NO.

AI-6-5394

5f. WORK UNIT NO.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 E.Glebe Rd
Alexandria, VA 22305

8. PERFORMING ORGANIZATION REPORT NO.

IDA Product 3001345

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OUSD(R&E)
Pentagon, Arlington, VA

10. SPONSOR'S / MONITOR'S ACRONYM(S)

OUSD(R&E)

11. SPONSOR'S / MONITOR'S REPORT NO(S).

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Maintaining Technology Advantage (MTA) Office within the Office of the Under Secretary of Defense, Research & Engineering focuses on protecting controlled technical information across the DoD from strategic competitors who may threaten U.S. military advantage. In support of MTA, IDA developed a decision aid for providing a high-level characterization of potential risk associated with strategic competitor acquisition of U.S. biological data. This product includes the decision aid and accompanying guidance for use.

15. SUBJECT TERMS

biological data; Biotechnology; bioeconomy; bio-cybersecurity; bioinformatics; genomics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NO. OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Patrick Lee
U	U	U	U	48	19b. TELEPHONE NUMBER (Include Area Code) (703) 571-4028

This page is intentionally left blank.