

IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems

IEEE Robotics and Automation Society

Developed by the
Standing Committee for Standards Activities

IEEE Std 7007™-2021

IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems

Developed by the

Standing Committee for Standards Activities
of the
IEEE Robotics and Automation Society

Approved 23 September 2021

IEEE SA Standards Board

Abstract: A set of ontologies with different abstraction levels that contain concepts, definitions, axioms, and use cases that assist in the development of ethically driven methodologies for the design of robots and automation systems is established by this standard. It focuses on the robotics and automation domain without considering any particular applications and can be used in multiple ways, for instance, during the development of robotics and automation systems as a guideline or as a reference “taxonomy” to enable clear and precise communication among members from different communities that include robotics and automation, ethics, and correlated areas. Users of this standard need to have a minimal knowledge of formal logics to understand the axiomatization expressed in Common Logic Interchange Format.

Keywords: artificial intelligence, automation, ethics, IEEE 7007™, ontology, robotics

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2021 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 12 November 2021. Printed in the United States of America.

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 978-1-5044-8013-0 STD24973
Print: ISBN 978-1-5044-8014-7 STDPD24973

IEEE prohibits discrimination, harassment, and bullying.

For more information, visit <https://www.ieee.org/about/corporate/governance/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.

Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (<https://standards.ieee.org/ipr/disclaimers.html>), appear in all standards and may be found under the heading “Important Notices and Disclaimers Concerning IEEE Standards Documents.”

Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Association (IEEE SA) Standards Board. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE Standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers are not necessarily members of IEEE or IEEE SA, and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE does not warrant or represent the accuracy or completeness of the material contained in its standards, and expressly disclaims all warranties (express, implied and statutory) not included in this or any other document relating to the standard, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; and quality, accuracy, effectiveness, currency, or completeness of material. In addition, IEEE disclaims any and all conditions relating to results and workmanlike effort. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE Standards documents are supplied “AS IS” and “WITH ALL FAULTS.”

Use of an IEEE standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document, should rely upon his or her own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Translations

The IEEE consensus development process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English version published by IEEE is the approved IEEE standard.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual shall not be considered or inferred to be the official position of IEEE or any of its committees and shall not be considered to be, nor be relied upon as, a formal position of IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter's views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group.

Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents.**

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and Standards Coordinating Committees are not able to provide an instant response to comments, or questions except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or in revisions to an IEEE standard is welcome to join the relevant IEEE working group. You can indicate interest in a working group using the Interests tab in the Manage Profile & Interests area of the [IEEE SA myProject system](#). An IEEE Account is needed to access the application.

Comments on standards should be submitted using the [Contact Us](#) form.

Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

Copyrights

IEEE draft and approved standards are copyrighted by IEEE under US and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These

include both use, by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, IEEE does not waive any rights in copyright to the documents.

Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; <https://www.copyright.com/>. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit [IEEE Xplore](#) or [contact IEEE](#). For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

Errata

Errata, if any, for all IEEE standards can be accessed on the [IEEE SA Website](#). Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in [IEEE Xplore](#). Users are encouraged to periodically check for errata.

Patents

IEEE Standards are developed in compliance with the [IEEE SA Patent Policy](#).

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at <https://standards.ieee.org/about/sasb/patcom/patents.html>. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting

inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

IMPORTANT NOTICE

IEEE Standards do not guarantee or ensure safety, security, health, or environmental protection, or ensure against interference with or from other devices or networks. IEEE Standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, and interference protection practices and all applicable laws and regulations.

Participants

At the time this IEEE standard was completed, the Ontologies for Ethically Driven Robotics and Automation Working Group had the following membership:

Edson Prestes, *Chair*
Sandro Rama Fiorini, *Vice Chair*
Mike Houghtaling, *Technical Editor*
Babita Ramlal, *Technical Editor*
Paulo Jorge Sequeira Gonçalves, *Secretary*

Bernd Blobel
Don Brutzman
Joel Carbonera
Abdelghani Chibani
Patrick Courtney
Nicola Fabiano
Tamas Haidegger

Vicky Hailey
Dennis Holstein
Tom Kurihara
Zvikomborero Murahwi
Joanna Isabelle Olszewska
Brian Page

William Remington Patterson
S. Veera Ragavan
Randy Rannow
Martin Saerbeck
André de O. Schenini Moreira
Ozlem Ulgen
Altaz Valani

The following members of the individual Standards Association balloting group voted on this standard. Balloters may have voted for approval, disapproval, or abstention.

Robert Aiello
M. Victoria Alonso
Pieter Botman
Diego Chiozzi
Murphy Choy
Jennifer Costley
Nicola Fabiano
Luis Andres Fajardo Arturo
Sandro Rama Fiorini
Paulo Goncalves
Didem Gurdur Broo
Tamas Haidegger
Robert Hobbs
Werner Hoelzl
Michael Houghtaling
Faiz Ikramulla

Piotr Karocki
Edmund Kienast
Robert Kozma
Loi Lei Lai
Sean Laroque-Doherty
Ting Li
Xiao Liang
Lars Luenenburger
Javier Luiso
Quintin McGrath
Lingzhong Meng
Lyria Bennett Moses
Rajesh Murthy
Alexander Novotny
Joanna Isabelle Olszewska
Brian Page

Sridhar Raghavan
Randy Rannow
Annette Reilly
John Sheppard
Matthew Silveira
Barnaby Simkin
Wayne Stec
Robert Soper
David Tepen
Ozlem Ulgen
Ionel Marius Vladan
Eleanor Watson
Stephen Webb
Robert Wortham
Naritoshi Yoshinaga
Yu Yuan

When the IEEE SA Standards Board approved this standard on 23 September 2021, it had the following membership:

Gary Hoffman, *Chair*
Jon Walter Rosdahl, *Vice Chair*
John D. Kulick, *Past Chair*
Konstantinos Karachalios, *Secretary*

Edward A. Addy
Doug Edwards
Ramy Ahmed Fathy
J. Travis Griffith
Thomas Koshy
Joseph L. Koepfinger*
David J. Law

Howard Li
Daozhuang Lin
Kevin Lu
Daleep C. Mohla
Chenhui Niu
Damir Novosel
Annette Reilly
Dorothy Stanley

Mehmet Ulema
Lei Wang
F. Keith Waters
Karl Weber
Sha Wei
Howard Wolfman
Daidi Zhong

*Member Emeritus

Introduction

This introduction is not part of IEEE Std 7007-2021, IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems.

Ontologies are formal specifications of a shared conceptualization denoting relevant concepts and relationships for a target domain of discourse. The set of concepts and relationships selected for formalization are those deemed appropriate for the type of ontology and its intended use. As a type of model, ontologies are abstractions of reality for the domain of interest and are comprised of classes, attributes, relationships, constraints, rules, and axioms. These components are used to express the shared commitments specified by the vocabulary of terms defined in the ontology. To enable viable ontology reuse, ontological models frequently utilize an architectural framework of three model levels: a top or foundational level, a middle (core) level, and a domain-specific application level. The ontology model for this standard is positioned as a core level specification. For more about ontology architectures, levels, and types see Aßman, Zschaler, and Wagner [B9] and Guizzardi [B25].¹ The use of ontologies for representing knowledge in any domain enables the following:

- Clear and formal definition of concepts for a domain
- Analysis of concepts and their relationships in searching of inconsistency, incompleteness, and redundancy
- Establishment of a language for use in the communication process among robotic and non-robotic systems from different manufacturers and among different stakeholders

The increasing complexity of robot design, human-robot interaction, and the increasing proliferation of robots in societies requires ontological standards to move beyond concepts and create unambiguous taxonomies with properties and examples of practical applications in defined use cases.

In 2015, the IEEE Robotics and Automation Society published its first standard: IEEE Std 1872™-2015, IEEE Standard Ontologies for Robotics and Automation. This standard established a series of ontologies about robotics and automation (R&A) to represent knowledge in R&A domain through a common set of terms and definitions that allows for explicit knowledge transfer among any group of humans, robots, and other artificial systems. Among these ontologies, Core Ontology for Robotics and Automation (CORA) was developed to be a high-level standard from which domain-specific efforts could emanate. CORA has generic concepts of the R&A domain (e.g., robot, robot group, and robotic system) to serve as basis for more focused ontologies in R&A to address specific information requirements.

IEEE Std 7007 complements IEEE Std 1872-2015 by focusing on the R&A domain, taking in into consideration the ethical dimension without being constrained to any application or kind of robot. In addition, IEEE Std 7007 complements the IEEE 7000 series of standards, such as the IEEE Std 7000™, IEEE Standard Model Process for Addressing Ethical Concerns during System Design, which focuses on ethical considerations at each phase of development to prevent negative or unintended consequences.

The creation of IEEE Std 7007 is timely, as the need for the ethical creation and use of R&A technologies has emerged as a vital aspect of advancing the well-being of human beings. Designers and manufacturers are increasingly giving robots the capability to interact and collaborate with humans, preferably in an ethical and safe manner to meet the needs of modern industry and societal advancements. As the complexity in design and new applications of R&A Systems arise, it is important for standards bodies to keep pace with advancements in the field in order to guide robot activities, support progress, and evaluate these for conformity and acceptability by society.

¹ The numbers in brackets correspond to those of the bibliography in Annex E.

In order to aid in the creation of value-sensitive design in R&A and align with what stakeholders (government, industry, academia, civil society) expect, in terms of benefits and a positive impact on human well-being, the consideration of applied ethics and a multidisciplinary approach to the standard development has been undertaken, with input from legal, philosophical, and engineering experts for this work. Thus, it is expected that IEEE Std 7007 can be used in multiple ways, for instance, during the development of R&A systems as a guideline or as a reference document to enable a clear and precise communication among members from different communities that include robotics and automation, ethics, and correlated areas of expertise.

Contents

1. Overview	11
1.1 Scope	11
1.2 Purpose	11
1.3 Word usage	12
2. Normative references.....	12
3. Definitions, acronyms, and abbreviations	12
3.1 Definitions	12
3.2 Acronyms and abbreviations	13
4. Ontologies for ethically aligned robotics and automation systems.....	13
4.1 Conventions	13
4.2 Background.....	14
4.3 Top-level definitions.....	15
4.4 ERAS top-level concepts.....	16
4.5 Norms and Ethical Principles	25
4.6 Data Privacy and Protection	40
4.7 Transparency and Accountability	57
4.8 Ethical Violation Management	69
Annex A (informative) Informative definitions.....	82
A.1 Top-level definitions.....	82
A.2 Norms and Ethical Principles.....	83
A.3 Data Protection and Privacy	88
A.4 Transparency and accountability	95
A.5 Ethical Violation Management	100
Annex B (informative) Ontology development.....	102
Annex C (informative) Use cases	104
C.1 Use Case Template	104
C.2 Norms and Ethics Use Case: Domestic Personal Assistant Robot.....	105
C.3 Ethical Violation Management Use Case: Data Privacy and Protection.....	107
C.4 Transparency use case: autonomous system behavior explanation.....	109
Annex D (informative) Distributed Responsibility Ascription for Autonomous Systems	112
Annex E (informative) Bibliography	114

IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems

1. Overview

1.1 Scope

This standard establishes a set of ontologies with different abstraction levels that contain concepts, definitions, axioms, and use cases that are deemed relevant and appropriate to establish ethically driven methodologies for the design of robots and automation (R&A) systems.

1.2 Purpose

The purpose of the standard is to establish a set of definitions and their relationships to enable the development of R&A in accordance with shared values and internationally accepted ethical principles that facilitate trust in the creation and use of R&A. Emphasis is placed on the alignment of ethics and engineering to enable communities to understand how to pragmatically design and implement these systems within the context of a values-based society. These definitions allow for precise communications among global experts of different domains that include robotics, automation, artificial intelligence (AI), and ethics.

The use of ontologies for representing knowledge in any domain has several benefits that include the following:

- a) A formal definition of concepts of a particular domain in a language-independent representation, i.e., they are not dependent on a specific programming language, however, they are formally described to be implemented in a target language
- b) Tools for analyzing concepts and their relationships in searching for inconsistency, incompleteness, and redundancy
- c) Language being used in the communication process among robots from different manufacturers

Users of this standard are responsible for being apprised of and referring to appropriate, applicable ethical criteria for consideration during system design. Moreover, users should consult all applicable laws and regulations. Conformance to the provisions of this voluntary standard does not constitute compliance to any applicable regulatory requirements. Users of the standard are responsible for observing or referring to the applicable laws and regulations.

1.3 Word usage

The word *shall* indicates mandatory requirements strictly to be followed in order to conform to the standard and from which no deviation is permitted (*shall* equals *is required to*).^{1,2}

The word *should* indicates that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required (*should* equals *is recommended that*).

The word *may* is used to indicate a course of action permissible within the limits of the standard (*may* equals *is permitted to*).

The word *can* is used for statements of possibility and capability, whether material, physical, or causal (*can* equals *is able to*).

2. Normative references

The following referenced documents are indispensable for the application of this document (i.e., they must be understood and used, so each referenced document is cited in text and its relationship to this document is explained). For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments or corrigenda) applies.

IEEE Std 1872™-2015, IEEE Standard Ontologies for Robotics and Automation.^{3,4}

ISO/IEC 24707:2018, Information technology—Common Logic (CL)—A framework for a family of logic-based languages.⁵

3. Definitions, acronyms, and abbreviations

3.1 Definitions

For the purposes of this document, the following terms and definitions apply. The IEEE Standards Dictionary Online should be consulted for terms not defined in this clause.⁶ Informative definitions are available in Annex A.

autonomous system: A system, either physically embodied or realized entirely within a software substrate, capable of performing tasks and behaviors with a high degree of autonomy making informed decisions without external direction and with the ability to adapt to changing conditions, knowledge, and constraints.

NOTE—Subclause 4.8 presents further context and motivation for the following three definitions reflecting various legal or technology-oriented preferences regarding norm violations and autonomous systems.

common world view: An ethical autonomous system domain analysis that adopts a middle ground between the Legal World View (LWV) and the Technology World View (TWV) by relying upon the concept of a maturity level of socio-technology governance capabilities to be achieved and certified for governments

¹ The use of the word *must* is deprecated and cannot be used when stating mandatory requirements, *must* is used only to describe unavoidable situations.

² The use of *will* is deprecated and cannot be used when stating mandatory requirements, *will* is only used in statements of fact.

³ IEEE publications are available from the Institute of Electrical and Electronics Engineers (<http://standards.ieee.org/>).

⁴ The IEEE standards or products referred to in Clause 2 are trademarks owned by the Institute of Electrical and Electronics Engineers, Incorporated.

⁵ ISO publications are available from the International Organization for Standardization (<http://www.iso.org/>) and the American National Standards Institute (<http://www.ansi.org/>).

⁶ *IEEE Standards Dictionary Online* is available at: <http://dictionary.ieee.org>. An IEEE Account is required for access to the dictionary, and one can be created at no charge on the dictionary sign-in page.

adopting ethically driven robotics and automation systems (ERAS) ontology commitments. The extent and type of responsibility ascriptions that can be ascribed to ethically aware autonomous systems would be based on the level of socio-technology governance achieved.

legal world view: An ethical autonomous system domain analysis predisposition asserting that current and foreseeable future legal systems do not and should not permit ascribing responsibility to autonomous systems for any norm violation, legal or ethical.

technology world view: An ethical autonomous system domain analysis predisposition asserting that emerging advances in artificial intelligence (AI) technology will soon motivate granting autonomous systems with formal and legal agency with consequential accountability and responsibility requirements.

3.2 Acronyms and abbreviations

AI	artificial intelligence
CLIF	Common Language Interchange Format
CORA	Core Ontology for Robotics and Automation
CWV	Common World View
DPP	Data Protection and Privacy
ERAS	ethically driven robotics and automation systems
EVM	Ethical Violation Management
LWV	Legal World View
NEP	Norms and Ethical Principles
R&A	robotics and automation
TA	Transparency and Accountability
TLO	top-level ontology
TWV	Technology World View
UML	Unified Modeling Language

4. Ontologies for ethically aligned robotics and automation systems

4.1 Conventions

This document presents the formal definitions for the development of ERAS using Common Logic Interchange Format (CLIF) notation, as defined in ISO/IEC 24707:2018⁷. To facilitate the understanding of the entire document, the ontologies developed were broken into a set of interrelated sub-ontologies. In several cases, the definition of a concept in one of the sub-ontologies is dependent upon a separate concept defined in a different sub-ontology. To make this cross reference clear, this standard uses the following

⁷ Information on references can be found in Clause 2.

notation X:Y to indicate the concept Y is defined at ontology X. For instance, ERAS-TLO:Method makes reference to the concept Method defined in Ontology ERAS-TLO.

This document also uses the shorthand notation proposed by Berardi, Calvanese, and De Giacomo [B11] and by Calvanese and De Giacomo in 2007 [B15] and 2020 [B16] to provide additional semantics to the axioms. As an example, consider the following axiom:

$$\text{(forall (x y) (if (relation x y) \\ \text{(and (EntityX x) \\ \text{(EntityY y))))})}$$

It asserts that if the relationship “relation” associated to two instances x and y exists, then x should be an instance of the entity EntityX while y should be an instance of entity EntityY. If the semantics of the relationship entails cardinality constraints such that instance x can be related to some number of p instances of the entity EntityY, then the following shorthand notation is used:

$$\text{(forall (x) (if (EntityX x) \\ \text{(>= p (\#\{ y | (and (relation x y) (EntityY y)) \}))}))}$$

where # is an operator that is used to indicate the number of elements of a particular set. In this case, # { y | (and (relation x y) (EntityY y)) } denotes the number of the elements of the set { y | (and (relation x y) (EntityY y)) } which is comprised of instances y from entity EntityY that participate in the relationship “relation” with the instance x.

Thus, the previous sentence indicates that the size of set { y | (and (relation x y) (EntityY y)) } should be greater than or equal to p. This is shorthand notation for the following equivalent CLIF expressions:

$$\text{(>= p (sizeof (setof y (and (relation x y) (EntityY y)))))}$$

where sizeof and setof are helper operator expressions defined in CLIF. The sizeof operator expression returns the number of elements in a set. The setof operator expression constructs a set of elements bound to a free variable y where each bound value satisfies the logical properties expressed as a conjunction of CLIF operator expressions. More information on this notation can be found in Berardi, Calvanese, and De Giacomo [B11], Calvanese and De Giacomo, 2007 [B15] and Calvanese and De Giacomo, 2020 [B16].

4.2 Background

Due to the complexity of the domain of interest, the Working Group decided to concentrate its efforts on four interrelated domains within the Ethically aligned Robotics and Automation Systems (ERAS) domain, as discussed in Annex B. These domains are as follows:

- *Norms and Ethical Principles (NEP)*: This subdomain formalizes aspects of ethical theories and principles that characterize the norms of expected behaviors for norm aware agents and autonomous systems.
- *Data Protection and Privacy (DPP)*: This subdomain formalizes relevant concepts and relationships characterizing the data protection and privacy rules and regulations that shall be observed and upheld by ethical agents and autonomous systems.
- *Transparency and Accountability (TA)*: This subdomain formalizes the concepts and relationships necessary to enable ethical autonomous systems with capabilities to provide informative explanations for plans and associated action.

- *Ethical Violation Management (EVM)*: This subdomain formalizes concepts and relationships associated with capabilities to detect, assess, and manage ethical violations in autonomous system behavior. In addition to ethical violation conceptualizations, this subdomain also addresses concepts and relationships governing accountability, responsibility, and legal notions of personhood for agents.

For each subdomain, a set of use cases was elaborated to represent relevant and realistic scenarios to help identify the conceptual and relationship terminology for the ontologies. Some examples of use cases can be found in Annex C. This process gave rise to formal models expressed through Unified Modeling Language (UML) diagrams that contain the main concepts and relationships for each subdomain together with a set of axioms written in CLIF to add semantics to the models. These models are presented and axiomatization is elaborated for each aforementioned subdomain in 4.4 through 4.8.

Partitioning the ERAS ontology into four subdomains was one strategy to manage the modeling complexity of the domain of interest. In addition, a graph transformation modeling technique similar to that proposed in Guizzardi, Figueiredo, Hedblom, and Poels [B26] was adopted. This approach utilizes a set of graph transformation rules to simplify fully articulated ontology models that have explicit formalizations for all conceptual categories in the target domain. The set of preconditions contained in each graph transformation rule enables the mapping of categories represented as classes into simpler properties with enumerated value types. Similar rules can be applied to reverse the mapping from enumerations back to categories as classes.

Similarly, this standard manages the complexity of the ERAS domain by representing potential points of extension for elaboration by utilizing category properties with enumerated data types for the range type of the respective properties. The property terms and related enumerated value terms in the vocabulary provide examples of category properties and value sets as candidate points of extension for target domain ontologies that reference the ERAS core ontology conceptualizations. To simplify the semantics of ERAS concepts and relationships, certain class properties are defined with enumerated data types that have informative definitions to denote their intended meaning within the context of their affiliated property class domain. This is as opposed to formalizing each of the enumerated values with individual category classes and axioms.

Concept property terms and associated terms denoting property data types are optional and depict example characterizations of the domain concepts for the respective property terms. In most cases, value terms listed in each enumerated data type have only the corresponding informative definitions to characterize the example semantics and do not possess further formalization in axioms. However, in a few cases, in order to fully express the meaning and commitments for the related domain category of a property, value terms from the property's enumerated data type are referenced in one or more axioms for the category concept. The axioms that formalize the meaning of a Norm derogation process are examples of this latter case. Specifically, these axioms specify the semantics of a Norm derogation process as an Agent Action and its effect on the state of a Norm by referencing value terms from the `norm_state` enumerated data type for the state property of the Norm category.

Target-domain ontologies that require more complex formalizations can apply relevant graph rewriting rules, either manually or automatically with supporting tools, to elaborate the ERAS core vocabulary. The set of informative definitions for the ERAS enumerated value terms are listed in Annex A together with the informative definitions of the concepts identified across all ontologies.

4.3 Top-level definitions

As a core ontology, the Ethically aligned Robotics and Automation Systems (ERAS) ontology represents a mid-level set of formalizations and commitments that are platform independent and intended to fit between an upper top-level or foundational ontology and lower domain and application specific ontologies. However, potential users of the standard may have different requirements regarding top-level ontology alignments. Some may require the commitments of one or more existing top-level ontologies while other

user communities would need only a minimal set of top-level commitments to complete the formalization of the concepts, terms, and commitments axiomatized in the ERAS ontology.

This standard introduces the ERAS top-level concepts that are intended to represent a standalone set of formalizations that complete the definitions and commitments expressed in the four ERAS subdomains. They define a minimal set of terms for that purpose. Users that do not require alignments with other existing foundational ontologies would use them.

4.4 ERAS top-level concepts

The ERAS top-level concepts and relationships are derived as a minimal set of ontological commitments appropriate for completing the formalization of concepts and relationships relevant to the characterization of ethically oriented agents and autonomous systems as identified in the four ERAS subdomains. While this subclause includes a complete set of axioms for the ERAS top-level concepts and axioms, the formalizations are not intended to be applicable as a top-level ontology in other contexts. Many of the concepts are similar to those found in other top-level ontologies such as SUMO (Niles and Pease [B39]), GFO (Herre [B27]), UFO (Guizzardi [B25]), and KR (Sowa [B50]). However, the set of concepts that comprise them is smaller than any of those taxonomies and all of the relevant axioms are expressed in CLIF. Figure 1 depicts the UML model for ERAS top-level concepts.

All individuals are entities:

(forall (x) (Entity x))

Entities are specialized into two subcategories that distinguish between Physical and Abstract conceptualizations.

(forall (x) (if (Physical x) (Entity x)))

(forall (x) (if (Abstract x) (Entity x)))

A Physical entity is an entity that has a location in space-time, i.e., physical entities are spatio-temporal located at a pose in the world at a specific time. Both Time and SpatioTemporalPlace are subcategories of Abstract, i.e.,

(forall (x) (if (SpatioTemporalPlace x) (Abstract x)))

(forall (x) (if (Time x) (Abstract x)))

(forall (x) (if (Physical x)
 (exists (p t)
 (and (SpatioTemporalPlace p)
 (Time t)
 (located_at x p)
 (present_at x t))))))

The two referenced relationships have the following formalizations:

(forall (d r) (if (located_at d r)
 (and (Physical d)
 (SpatioTemporalPlace r))))

(forall (d r) (if (present_at d r)
 (and (Physical d)
 (Time r))))

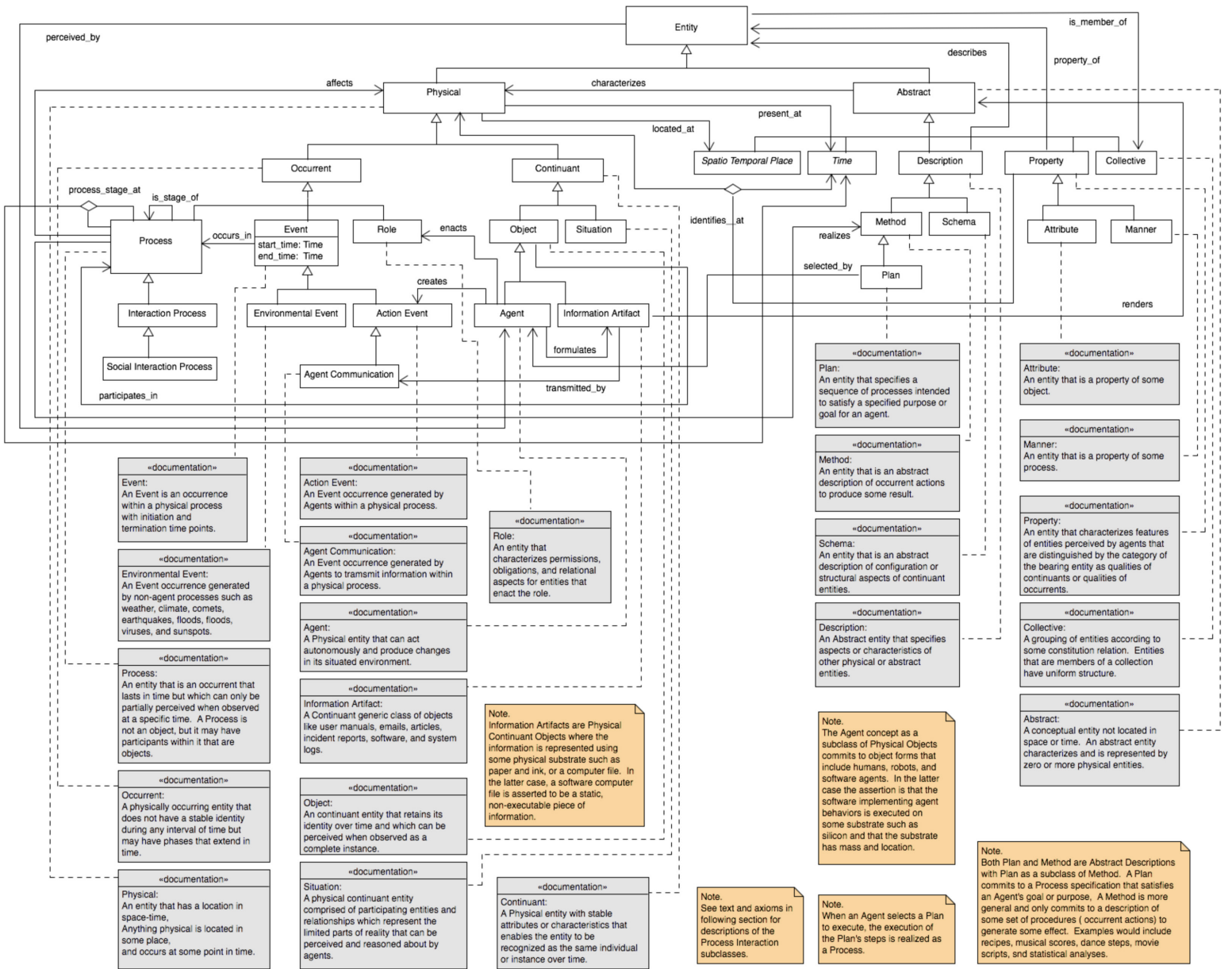


Figure 1 — Ethically driven robotics and automation systems top-level concepts UML diagram

The Abstract category classifies entities that are conceptual only and not located in space or time.

```
(forall (a) (if (Abstract a)
              (and (Entity a)
                   (not (Physical a))))))
```

```
(forall (p) (if (Physical p)
              (and (Entity p)
                   (not (Abstract p))))))
```

Abstract entities characterize and represent zero or more Physical entities.

```
(forall (x y) (if (characterizes x y)
                 (and (Abstract x)
                     (Physical y))))
```

The cardinality of this relationship is represented using shorthand notation for expressing multiplicity constraints on binary associations.

```
(forall (a) (if (Abstract a)
              (>= 0 ( # { p | (and (characterizes a p) (Physical p) ) } ))))
```

The Physical category is specialized into two subcategories that distinguish between Occurrent and Continuant conceptualizations.

```
(forall (x) (if (Occurrent x) (Physical x)))
```

```
(forall (x) (if (Continuant x) (Physical x)))
```

The Occurrent category classifies physically occurring entities that do not have a stable identity during any interval of time. Occurrent entities happen in time and may have temporal parts or phases that extend in time but cannot be wholly perceived at any point in time. Occurrent entities are in states of flux that prevent them from being recognized by a stable set of properties.

```
(forall (o) (if (Occurrent o)
              (exists(t1 t2)
                (and (Time t1)
                    (Time t2)
                    (not (= t1 t2))
                    (present_at o t1)
                    (present_at o t2)
                    (not (exists a)
                      (and (Property a)
                          (identifies_at a o t1)
                          (identifies_at a o t2))))))))
```

```
(forall (a p t) (if (identifies_at a p t)
                   (and (Property a)
                       (Physical p)
                       (Time t))))
```

The Occurrent category is specialized into three subcategories to distinguish between concepts of Process, Event, and Role.

```
(forall (x) (if (Process x)(Occurrent x)))
```

(forall (x) (if (Event x) (Occurrent x)))

(forall (x) (if (Role x) (Occurrent x)))

A Process entity is an Occurrent entity that lasts in time but that can only be partially perceived when observed at a specified time. A Process is not an object but may have participants within it that are objects, and it may be perceived in stages, at least one, as it occurs over time.

(forall (p) (if (Process p)
 (exists (t q a)
 (and (Agent a)
 (Time t)
 (Process q)
 (process_stage_at q p t)
 (not (perceived_by p a))
 (perceived_by q a))))))

(forall (s p t) (if (process_stage_at s p t)
 (and (Process s)
 (Process p)
 (is_stage_of s p)
 (Time t))))

(forall (d r) (if (is_stage_of d r)
 (and (Process d)
 (Process r))))

(forall (d r) (if (perceived_by d r)
 (and (Entity d)
 (Agent r))))

A Process may affect a Physical entity.

(forall (p e) (if (affects p e)
 (and (Process p)
 (Physical e))))

A Role entity is an Occurrent entity that characterizes permissions, obligations, and relational aspects for Agents that enact the Role.

(forall (a r) (if (enacts a r)
 (and (Agent a)
 (Role r))))

An Event entity is an Occurrent entity within a physical Process with initiation and termination time points.

(forall (e) (if (Event e)
 (exists (t n p)
 (and (Process p)
 (Time t)
 (Time n)
 (start_time e t)
 (end_time e n)
 (occurs_in e p))))))

(forall (e p) (if (occurs_in e p)

(and (Event e)
(Process p))))

(forall (e t) (if (or (start_time e t)
(end_time e t)
(and (Event e)
(Time t))))))

The Event category is specialized into two subcategories: Action Event and Environmental Event.

(forall (x) (if (ActionEvent x) (Event x)))

(forall (x) (if (EnvironmentalEvent x) (Event x)))

(forall (e) (if (EnvironmentalEvent e)
(not (ActionEvent e))))

(forall (e) (if (ActionEvent e)
(not (EnvironmentalEvent e))))

An Action Event entity is an Event occurrence created by Agents within a process.

(forall (e) (if (ActionEvent e)
(exists (a p)
(and (Agent a)
(Process p)
(occurs_in e p)
(creates a e))))))

(forall (a e) (if (creates a e)
(and (Agent a)
(ActionEvent e))))

An Environmental Event entity is an Event occurrence generated by non-agent processes.

(forall (e) (if (EnvironmentalEvent e)
(exists (p)
(and (Process p)
(occurs_in e p)
(not (and (InteractionProcess p)
(ActionEvent e))))))

The Action Event category is specialized with the Agent Communication subcategory.

(forall (x) (if (AgentCommunication x) (ActionEvent x)))

An Agent Communication entity is an Action Event generated by Agents to transmit information within a Physical Process.

(forall (x) (if (AgentCommunication x)
(exists (d)
(and (InformationArtifact d)
(transmitted_by d x))))))

(forall (d r) (if (transmitted_by d r)

(and (InformationArtifact d)
(AgentCommunication r))))

The Process category has Interaction Process as a subcategory specialization.

(forall (i) (if (InteractionProcess i) (Process i)))

An Interaction Process is a Process that includes at least one Action Event that was generated by one or more Agents.

(forall (i) (if (InteractionProcess i)
(exists (a e)
(and (Agent a)
(ActionEvent e)
(creates a e)
(occurs_in e i)
(participates_in a i))))))

(forall (d r) (if (participates_in d r)
(and (Object d)
(Process r))))

The Interaction Process category has Social Interaction Process as a subcategory specialization.

(forall (s) (if (SocialInteractionProcess s) (InteractionProcess s)))

A Social Interaction Process is an Interaction Process that includes multiple Agents engaged in Agent Communication sub process stages.

(forall (s) (if (SocialInteractionProcess s)
(exists (a b c)
(and (Agent a)
(Agent b)
(not (= a b))
(AgentCommunication c)
(participates_in a s)
(participates_in b s)
(occurs_in c s))))))

The Continuant category classifies physical entities with stable attributes or characteristics that enable the entity to be recognized as the same individual or instance over time. This category is specialized into subcategories to distinguish the Object and Situation conceptualizations.

(forall (x) (if (Object x) (Continuant x)))

(forall (x) (if (Situation x) (Continuant x)))

(forall (c) (if (Continuant c)
(exists (a t1 t2)
(and (Attribute a)
(Time t1)
(Time t2)
(not (= t1 t2))
(present_at c t1)
(identifies_at a c t1))))

(present_at c t2)
(identifies_at a c t2))))))

An Object entity is a Continuant entity that retains its identity over time and which can be perceived when observed as a complete instance. Object entities can have different properties at different times and therefore can undergo change.

(forall (o) (if (Object o)
(and (Continuant o)
(not (Situation o))))))

The Situation category is a physical Continuant entity comprised of participating entities and relationships that represent the limited parts of reality that can be perceived and reasoned about by agents.

(forall (s) (if (Situation s)
(and (Continuant s)
(not (Object s))))))

The Object category is specialized into the subcategories of Agent and Information Artifact.

(forall (x) (if (Agent x) (Object x)))

(forall (x) (if (InformationArtifact x) (Object x)))

An Agent entity is a Physical entity that can create Action Events to produce changes in the agent's physical situation and which can be perceived by it or other agents. An Agent has a repertoire of Plans that describe Methods for achieving or satisfying agent intentions. The plan methods specify sequences of Process actions that affect changes in the Physical situation of the Agent.

(forall (d r) (if (selected_by d r)
(and (Plan d)
(Agent r))))

(forall (a p m e) (if (and (Agent a)
(Process p)
(participates_in a p)
(Plan m)
(Physical e)
(not (= e p))
(realizes p m)
(affects p e))
(selected_by m a)))

The Information Artifact category is a Continuant generic class of Objects that renders abstract descriptive ideas, expressions, and facts as tangible artifacts using printed text, electronic media or some form of physical substrate. Information Artifact entities are formulated by Agents. Examples in the ERAS domain include user manuals, system logs, incident reports, articles, emails, and software.

(forall (d r) (if (formulates d r)
(and (Agent d)
(InformationArtifact r))))

(forall (d r) (if (renders d r)
(and (InformationArtifact d)

(Abstract r)))

(forall (x) (if (InformationArtifact x)
 (exists (a)
 (and (Agent a)
 (formulates a x))))))

The Abstract category is specialized into three instantiatable subcategories: Description, Property, and Collective, and two abstract, non-instantiatable subcategories: Time and Spatio Temporal Place.

(forall (x) (if (Description x) (Abstract x)))

(forall (x) (if (Property x) (Abstract x)))

(forall (x) (if (Collective x) (Abstract x)))

(forall (x) (if (Time x) (Abstract x)))

(forall (x) (if (SpatioTemporalPlace x) (Abstract x)))

The non-instantiatable abstract Time category represents a linear sequence of time points. The non-instantiatable abstract Spatio Temporal Place category represents the combinatorial properties of space and time qualities that characterize locations of Occurrent and Continuant Entities. Physical entities are present at or exist at various points in Time and are located at various Spatio Temporal Places.

The Description category classifies entities that specify aspects or characteristics of other physical or abstract entities.

(forall (d) (if (Description d)
 (exists (e)
 (and (Entity e)
 (describes d e))))))

(forall (d e) (if (describes d e)
 (and (Description d)
 (Entity e))))

The Description category is specialized into two subcategories: Method and Schema.

(forall (x) (if (Method x) (Description x)))

(forall (x) (if (Schema x) (Description x)))

The Method category classifies an entity that is an abstract description of Occurrent Process actions to produce some result.

(forall (x) (if (Method x)
 (exists (y)
 (and (Process y)
 (describes x y)
 (realizes y x))))))

(forall (d r) (if (realizes d r)
 (and (Process d)
 (Method r))))

The Method category is specialized into the Plan subcategory.

(forall (x) (if (Plan x) (Method x)))

The Plan category classifies an entity that specifies a sequence of processes intended to satisfy a specified purpose or goal for an Agent by affecting changes in the Agent's Physical situation.

(forall (x p) (if (and (Plan x)
 (Process p)
 (realizes p x))
 (exists (e)
 (and (Physical e)
 (not (= e p))
 (affects p e))))))

The Schema category classifies an entity that is an abstract description of configuration or structural aspects of Continuant entities.

(forall (x) (if (Schema x)
 (exists (c)
 (and (Continuant c)
 (describes x c))))))

The Property category classifies an entity that characterizes features of entities perceived by Agents that are distinguished by the category of the bearing entity as qualities of Continuants or qualities of Occurrents.

(forall (p) (if (Property p)
 (exists (a e)
 (and (Agent a)
 (Entity e)
 (property_of p e)
 (perceived_by e a))))))

The Property category is specialized into two subcategories: Attribute and Manner.

(forall (x) (if (Attribute x) (Property x)))

(forall (x) (if (Manner x) (Property x)))

The Attribute category classifies an entity that is a property of some object, i.e.,

(forall (x) (if (Attribute x)
 (exists (o)
 (and (Object o)
 (property_of x o))))))

The Manner category classifies an entity that is a property of some process, i.e.,

(forall (x) (if (Manner x)
 (exists (p)
 (and (Process p)
 (property_of x p))))))

(forall (d r) (if (property_of d r)
 (or (and (Attribute d)

(Object r)
(and (Manner d)
(Process r))))

The Collective category classifies entities that are grouped together according to some constitution relation. Entities that are members of a collection have uniform structure.

(forall (d r) (if (is_member_of d r)
(and (Entity d)
(Collective r))))

(forall (c) (if (Collective c)
(exists (e r)
(and (Entity e)
(is_member_of e c)
(Description r)
(describes r c))))))

4.5 Norms and Ethical Principles

Figure 2 shows the UML diagram that captures the main concepts and relationships elicited during the investigation of the Norms and Ethical Principles (NEP) subdomain. It focuses on aspects of ethical theories and principles that characterize the norms of expected behaviors for norm-oriented agents and autonomous systems. This diagram links some concepts to ones already represented in the ERAS top-level ontology and other related ontologies.

A Norm is a Method entity that describes a set of rules and methods governing behavior expected for norm-aware agents. Norm types are derived from respective ethical theories and are possibly influenced by agent social contexts. An ethical theory is a systematization of concepts specifying or recommending aspects of morally correct behavior based on philosophical values and the characterization of right and wrong conduct. For norm aware agents, normative ethical theory is concerned with the practical means of determining a moral course of action.

(forall (n) (if (Norm n) (ERAS-TLO:Method n)))

(forall (t) (if (EthicalTheory t) (ERAS-TLO:Method t)))

(forall (n) (if (Norm n)
(exists (t s)
(and (EthicalTheory t)
(SocialCollection s)
(specifies_norm_modality t n)
(is_prescribed_by n t)
(influences_norm_applicability s n))))))

Note:
Green classes: ERAS-TLO Top Level Ontology
White classes: ERAS-NEP Norms & Ethics Principles

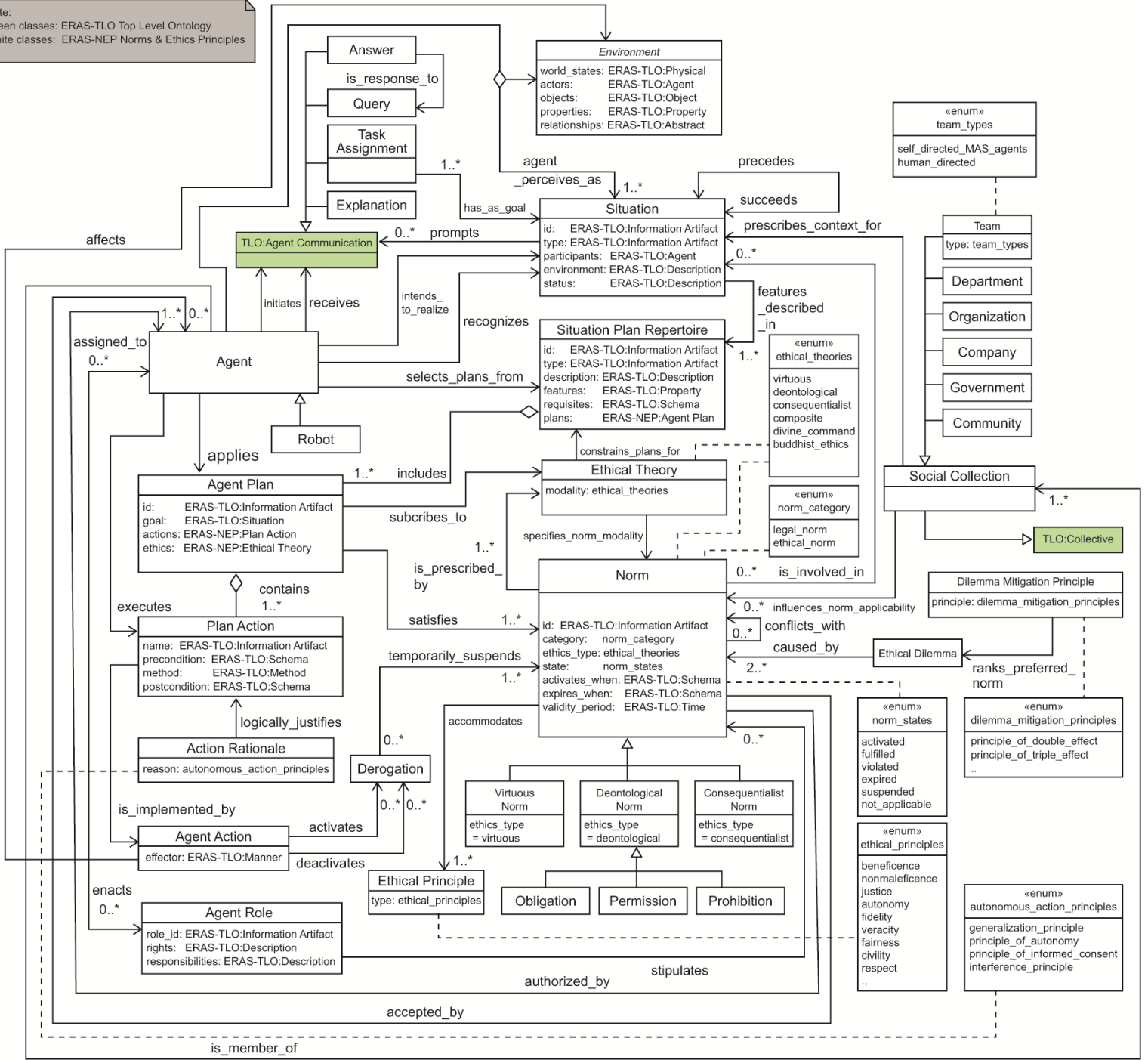


Figure 2 — Norms and Ethical Principles UML diagram

A descriptive characterization of representative forms of Ethical Theories includes the following set of enumerated examples.

```
( = ( Description ethical_theories)
  { virtuous
    deontological
    consequentialist
    composite
    divine_command
    Buddhist_ethics } )
```

The various philosophical forms of Ethical Theories aim to specify the modality of Norms intended to guide the behaviors of ethically aware Agents. Each Norm is prescribed by at least one Ethical Theory.

```
(forall (d r) (if (specifies_norm_modality d r)
  (and (EthicalTheory d)
    (Norm r) )))
```

```
(forall (d r) (if (is_prescribed_by d r)
  (and (Norm d)
    (EthicalTheory r) )))
```

Using the shorthand notation,

```
(forall (n) (if (Norm n)
  (>= 1 ( # { t | (and (is_prescribed_by n t) (EthicalTheory t) } ) )))
```

An extensive set of Norms exists for Ethically-aware Agents to consider. For autonomous agents, Norms are authorized by one or more Agents.

```
(forall(d) (if (Norm d)
  (exists(r)
    (and (Agent r)
      (authorized_by d r))))))
```

```
(forall (n a) (if (authorized_by d r)
  (and (Norm d)
    (Agent r))))
```

Using shorthand notation,

```
(forall (n) (if (Norm n)
  (>= 1 ( # { a | (and (authorized_by n r) (Agent a) } ) )))
```

Agents may accept zero or more Norms.

```
(forall (d r) (if (accepted_by d r)
  (and (Norm d)
    (Agent r))))
```

(forall (n) (if (Norm n)
(>= 0 (# { a | (and (accepted_by n a) (Agent a) })))))

The Norm category has at least three subcategories representing distinct behavioral courses of actions expressed by the main ethical theories.

(forall (n) ((if (VirtuousNorm n) (Norm n)))

(forall (n) ((if (DeontologicalNorm n) (Norm n)))

(forall (n) ((if (ConsequentialistNorm n) (Norm n)))

A Virtuous Norm is derived from the virtuous ethical theory that elucidates correct action choices based on alignment with certain dispositional character traits or virtues that are appropriate and praiseworthy. From this perspective, correct agent behavior is achieved by adhering to character traits deemed praiseworthy and not blameworthy.

(forall (n) (if (VirtuousNorm n) (= (ethics_type n) virtuous)))

A Deontological Norm is derived from the deontological ethical theory that stipulates correct action choices based on the action's conformity to universal rules for judging rightness or wrongness of an act. From this perspective, correct behavior is independent of the resulting consequences.

(forall (n) (if (DeontologicalNorm n) (= (ethics_type n) deontological)))

A Consequentialist Norm is derived from the Consequentialist ethical theory that elucidates correct action choices based on the consequences that the action produces. Generally, actions that are expected to result in a greater intrinsic good are preferred.

(forall (n) (if (ConsequentialistNorm n) (= (ethics_type n) consequentialist)))

Additionally, Deontological Norms can be further classified as follows:

— *Obligation*: What an agent should do — an attribute that applies to propositions that an agent is required by some authority to make true,

(forall (n) (if (Obligation n) (DeontologicalNorm n))).

— *Permission*: What an agent may do — an attribute that applies to propositions that an agent is permitted, by some authority to make true,

(forall (n) (if (Permission n) (DeontologicalNorm n))).

— *Prohibitions*: What an agent is forbidden to do — an attribute that applies to propositions that an agent is forbidden, by some authority to make true,

(forall (n) (if (Prohibition n) (DeontologicalNorm n))).

Some Norms are specified and enforced as laws within a social legal system. Consequently, the ERAS ontology commits to an enumerated norm type to distinguish between legal norm and ethical norm categories.

(= (Description norm_category)
{ legal_norm
ethical_norm })

Norm methods influence the behaviors of norm-aware Agents by constraining the Agent Plans that Agents select for realizing Agent objectives and goals. In this context, Norm methods have life cycle states. A descriptive characterization of these states includes the following enumerated examples.

```
( = ( Description norm_states )  
  { activated  
    fulfilled  
    violated  
    expired  
    suspended  
    not_applicable } )
```

Norm methods accommodate ethical principles that specify descriptions of general moral proposition and value judgments that characterize and justify particular ethical prescriptions and evaluations of agent actions. The ERAS ontology commits to having each Norm specification accommodate at least one Ethical Principle.

```
(forall (p) (if (EthicalPrinciple p) (ERAS-TLO:Method p)))
```

```
(forall (d r) (if (accommodates d r)  
  (and (Norm d)  
    (EthicalPrinciple r))))
```

```
(forall (n) (if (Norm n)  
  (exists (p)  
    (and (EthicalPrinciple p)  
      (accommodates n p) )))))
```

A descriptive characterization of representative ethical principles includes the following enumerated examples:

```
( = ( Description ethical_principles )  
  { beneficence  
    nonmaleficence  
    justice  
    autonomy  
    fidelity  
    veracity  
    fairness  
    civility  
    respect } )
```

Agent Plans typically need to satisfy more than one Norm and consequently the multiple constraints represented by the requisite Norms may lead to conflicts or Ethical Dilemmas.

An Ethical Dilemma is a decision-making situation arising between conflicting normative rules of behavior in which none of the choices are deemed unambiguously acceptable or preferable. Ethical Dilemmas are caused by, at least, two such conflicting Norms.

```
(forall (d) (if (EthicalDilemma d) (ERAS-TLO:Situation d)))
```

```
(forall (d) (if (EthicalDilemma d)
```

```
(exists (a b)
  (and (Norm a)
    (Norm b)
    (not (= a b))
    (caused_by d a)
    (caused_by d b)
    (or (conflicts_with a b) (conflicts_with b a))))))
```

```
(forall (d r) (if (conflicts_with d r)
  (and (Norm d)
    (Norm r)
    (not (= d r))))))
```

A Norm may conflict with zero or more norms. Using shorthand notation,

```
(forall (n) (if (Norm n)
  (>= 0 ( # { c | (and (conflicts_with n c) (Norm c) ) } ))))
```

A Norm involved in a Norm conflict may be one of the norms causing an Ethical Dilemma.

```
(forall (d r) (if (caused_by d r)
  (and (EthicalDilemma d)
    (Norm r)
  )))
```

It takes at least two norms to cause an Ethical Dilemma.

```
(forall (d) (if (EthicalDilemma d)
  (>= 2 ( # { n | (and (caused_by d n) (Norm n) ) } ))))
```

Ethical Dilemmas can be resolved by applying various Dilemma Mitigation Principles to rank a preferred Norm among the conflicting Norms.

```
(forall (p) (if (DilemmaMitigationPrinciple p) (ERAS-TLO:Method p)))
```

```
(forall (d r) (if (ranks_preferred_norm d r)
  (and (DilemmaMitigationPrinciple d)
    (EthicalDilemma r)
  )))
```

A descriptive characterization of representative dilemma mitigation principles includes the following enumerated examples:

```
(= ( Description dilemma_mitigation_principles)
  { principle_of_double_effect
    principle_of_triple_effect } )
```

Norms impact how Agents act in their situated environment. The ERAS Agent category is defined in the ERAS Top-Level Concepts ontology as a subcategory of a Physical Continuant Object. An Agent entity is an entity that can act on its own and produce changes in its situated environment. The NEP subdomain specializes the ERAS Top-Level Concept with additional relationships.

(forall (x) (if (Agent x) (ERAS-TLO:Agent x)))

The Agent category is specialized with a Robot subcategory, which is intended to be conceptually equivalent to the CORA:Robot concept without requiring the direct importation of the CORA ontology.

(forall (x) (iff (Robot x) (CORA:Robot x)))

Agents perceive and act in an environment. In the context of the ERAS ontology, the Environment category is an abstraction that classifies an external collection of entities, entity properties, entity relationships, and occurrent processes that pose potential internal Agent conceptualizations derived from Agent perceptions of the external entities present in the Environment. The set of entity conceptualizations that are internalized as Agent perceptions is bounded by the type, capabilities, and focus of the sensors employed by the Agent, and consequently may be a subset of the entities present in the environment. This limited entity set represents that portion of an Agent's environmental reality about which it can perceive, interpret, and reason. Note however that the entities present in the Environment are not dependent upon Agent perceptions. Internally, an Agent's perceptions of its external circumstances are represented and classified by the Situation category.

As autonomous actors, Norm-aware Agents perceive, recognize, and become aware of Situations presented in their environments. They respond with the selection and application of appropriate Agent Plans for realizing goals and reacting to recognized Situations.

This ontology subdomain extends the Situation category as described in the ERAS Top-Level Concepts with additional relationships between the Agent and Environment categories.

(forall (e) (if (Environment e) (ERAS-TLO:Continuant e)))

(forall (s) (if (Situation s) (ERAS-TLO:Situation s)))

(forall (s) (if (Situation s)
 (exists (e a)
 (and (Environment e)
 (Agent a)
 (agent_perceives_as a e s)
 (recognizes a s))))))

(forall (a e s) (if (agent_perceives_as a e s)
 (and (Agent a)
 (Environment e)
 (Situation s))))

(forall (d r) (if (recognizes d r)
 (and (Agent d)
 (Situation r))))

(forall (d r) (if (intends_to_realize d r)
 (and (Agent d)
 (Situation r))))

For any environment in which an agent is a participant, the agent formulates at least one situation representation that is its perception of the environmental state. Using shorthand notation,

(forall (a e) (if (and (Environment e)
 (Agent a))
 (>= 1 (#{ s | (and (agent_perceives_as a e s) (Situation s)) })))))

Situations, as Agent perceptions of environmental circumstances, may involve zero or more Norms.

(forall (d r) (if (is_involved_in d r)
 (and (Norm d)
 (Situation r))))

(forall (n) (if (Norm n)
 (>= 0 (#{ s | (and (is_involved_in n s) (Situation s)) })))))

As Agents interact with their Environment and sense the world states represented as Situations, their perceptions and intentions generate Situation sequences establishing succeeds and precedes relationships.

(forall (d r) (if (or (succeeds d r)
 (precedes d r))
 (and (Situation d)
 (Situation r)
 (not (= d r)))))

Agents interact with their environment by selecting Agent Plans from a repertoire of plans that are relevant for the goals and responses determined by Agent reasoning processes. Agent Plans that are included in a Situation Plan Repertoire consist of specifications, partial or complete, for a sequence of agent actions to achieve target goals, objectives, and services to realize agent intentions. Agent Plans as subclasses of ERAS-TLO:InformationArtifact render ERAS-TLO:Abstract Plans into some physical substrate.

(forall (p) (if (AgentPlan p) (ERAS-TLO:InformationArtifact p)))

(forall (x) (if (AgentPlan x)
 (exists (p)
 (and (ERAS-TLO:Plan p)
 (renders x p)))))

(forall (r) (if (SituationPlanRepertoire r) (ERAS-TLO:InformationArtifact r)))

(forall (p) (if (AgentPlan p)
 (exists (r t pa aa n)
 (and (SituationPlanRepertoire r)
 (EthicalTheory t)
 (PlanAction pa)
 (AgentAction aa)
 (Norm n)
 (includes r p)
 (subscribes_to p t)
 (constrains_plans_for t r)
 (contains p pa)
 (is_implemented_by pa aa)
 (satisfies p n)))))

(forall (d r) (if (includes d r)

(and (SituationPlanRepertoire d)
(AgentPlan r)))

A Situation Plan Repertoire includes at least one Agent Plan. Using shorthand notation,

(forall (r) (if (SituationPlanRepertoire r)
(≥ 1 (#{ p | (and (includes r p) (AgentPlan p)) }))))

Agent Plans included in the Situation Plan Repertoire are constrained by Ethical Theories. And each Agent Plan satisfies one or more Norms that are prescribed by the associated Ethical Theory.

(forall (d r) (if (constrains_plans_for d r)
(and (EthicalTheory d)
(SituationPlanRepertoire r))))

(forall (d r) (if (satisfies d r)
(and (AgentPlan d)
(Norm r))))

(forall (p) (if (AgentPlan p)
(≥ 1 (#{ n | (and (satisfies p n) (Norm n)) }))))

(forall (d r) (if (subscribes_to d r)
(and (AgentPlan d)
(EthicalTheory r))))

The commitments of this ontology subdomain asserts that for any Situation that is an Agent's perception of its environment, there exists at least one set of feature descriptions in the Situation Plan Repertoire that characterizes the situation.

(forall (d r) (if (features_described_in d r)
(and (Situation d)
(SituationPlanRepertoire r))
))

(forall (r) (if (SituationPlanRepertoire r)
(≥ 1 (#{ s | (and (features_described_in s r) (Situation s)) }))))

This enables an Agent to select relevant Agent Plans from its Situation Plan Repertoire that are qualitatively good matches for circumstances in the current environment situation.

(forall (r a s) (if (and (SituationPlanRepertoire r)
(Agent a)
(Situation s)
(features_described_in s r)
(or (recognizes a s)
(intends_to_realize a s)))
(selects_plans_from a r)))

(forall (d r) (if (selects_plans_from d r)
(and (Agent d)

(SituationPlanRepertoire r))))

(forall (d r) (if (applies d r)
 (and (Agent d)
 (AgentPlan r))))

Agent Plans contain one or more Plan Actions. A Plan Action, as a constituent of an Agent Plan, specifies the preconditions and postconditions for the application of an agent action to achieve the objectives and goals of the plan. Plan Actions as subclasses of ERAS-TLO:InformationArtifact render ERAS-TLO:Abstract Methods into some physical substrate.

(forall (x) (if (PlanAction x) (ERAS-TLO:InformationArtifact x)))

(forall (x) (if (PlanAction x)
 (exists (m)
 (and (ERAS-TLO:Method m)
 (renders x m))))))

(forall (d r) (if (contains d r)
 (and (AgentPlan d)
 (PlanAction r))))

(forall (p) (if (AgentPlan p)
 (>= 1 (# { a | (and (contains p a) (PlanAction a)) }))))

A Plan Action may have a supporting Action Rationale that logically justifies it as an autonomous action.

(forall (r) (if (ActionRationale r) (ERAS-TLO:Method r)))

(forall (d r) (if (logically_justifies d r)
 (and (ActionRationale d)
 (PlanAction r))))

Reasons for the justification of a Plan Action are based on autonomous action principles (see Hooker and Kim [B29]) similar to conventional principles of reciprocity. A descriptive characterization of such principles includes the following enumerated examples:

(= (Description autonomous_action_principles)
 { generalization_principle
 principle_of_autonomy
 principle_of_informed_consent
 interference_principle })

When Plan Action preconditions hold, the Agent applying the Agent Plan of the constituent Plan Action executes the Plan Action.

(forall (d r) (if (executes d r)
 (and (Agent d)
 (PlanAction r))))

Plan Actions are implemented by Agent Actions. An Agent Action is an operation or effector that implements the method specified in the Plan Action. The Agent Action is applied and executed by the Agent to affect state changes in an Agent's situated environment.

(forall (x) (if (AgentAction x) (ERAS-TLO:Process x)))

(forall (d r) (if (is_implemented_by d r)
(and (PlanAction d)
(AgentAction r))))

(forall (a pa aa) (if (and (Agent a)
(PlanAction pa)
(AgentAction aa)
(is_implemented_by pa aa)
(executes a pa))
(exists (e)
(and (Environment e)
(affects aa e))))))

While applying Agent Plans that satisfy the set of Norms to which the Agent Plan subscribes, Ethical-aware Agents may need to temporarily suspend or derogate a Norm instance. The Derogation category is the Process that an Agent activates for that purpose. When a Norm instance can be resumed, the Agent deactivates the Derogation.

(forall (d) (if (Derogation d) (ERAS-TLO:Process d)))

(forall (d r) (if (temporarily_suspends d r)
(and (Derogation d)
(Norm r))))

(forall (d r) (if (or (activates d r)
(deactivates d r))
(and (AgentAction d)
(Derogation r))))

(forall (d a) (if (and (Derogation d)
(AgentAction a)
(activates a d))
(exists (n)
(and (Norm n)
(temporarily_suspends d n)
(= (state n) suspended))))))

(forall (d a n) (if (and (Norm n)
(Derogation d)
(AgentAction a)
(activates a d)
(temporarily_suspends d n))
(= (state n) suspended))))

```
(forall (d a n) (if (and (Norm n)
    (= (state n) suspended)
    (Derogation d)
    (AgentAction a)
    (deactivates a d))
    (= (state n) activated))))
```

Agents formulate and represent their objective goals and responses as Situations that they intend to realize by selecting an Agent Plan that affects their situated environment for the purpose of achieving the intended objectives.

```
(forall (a ap pa aa e s) (if (and (Agent a)
    (AgentPlan ap)
    (PlanAction pa)
    (AgentAction aa)
    (Environment e)
    (Situation s)
    (contains ap pa)
    (is_implemented_by pa aa)
    (agent_perceives_as a e s)
    (or (recognizes a s)
        (intends_to_realize a s))
    (affects aa e))
    (and (applies a ap)
        (executes a pa))))
```

Any agent can interact with other agents by initiating and receiving Agent Communication Action Events to transmit and exchange Information Artifacts. The Agent Communication category defined by the ERAS Top-Level Ontology (TLO) is an Action Event subcategory. It is specialized in the NEP subdomain with additional relationships and with four subcategories.

```
(forall (d r) (if (initiates d r)
    (and (Agent d)
        (ERAS-TLO:AgentCommunication r))))
```

```
(forall (d r) (if (receives d r)
    (and (Agent d)
        (ERAS-TLO:AgentCommunication r))))
```

```
(forall (c) (if (ERAS-TLO:AgentCommunication c)
    (exists (a i)
        (and (Agent a)
            (ERAS-TLO:InformationArtifact i)
            (or (initiates a c)
                (receives a c))
            (transmitted_by i c)
        )))
```

```
(forall (x) (if (Explanation x) (ERAS-TLO:AgentCommunication x)))
```

(forall (x) (if (TaskAssignment x) (ERAS-TLO:AgentCommunication x)))

(forall (x) (if (Query x) (ERAS-TLO:AgentCommunication x)))

(forall (x) (if (Answer x) (ERAS-TLO:AgentCommunication x)))

The Agent Communication subcategories are defined informally as follows:

- *Explanation*: A response to a request to explain and justify system behavior. The response may be tailored to the type and role of the agent making the request.
- *Task Assignment*: A communication that assigns and specifies a mission, chore, duty, problem, or goal to undertake and accomplish or solve. The task specification may include initial conditions, a goal, assertions, and characterizations of available operations and resources, which are then represented in a task goal situation.
- *Query*: A communication requesting information from some source about some topic. The inquiry may be expressed informally in natural language, formally using some formal query language, or using some visual medium.
- *Answer*: A communication responding with information that answers prior queries. The response may be expressed informally in natural language, formally using some formal query language, or using some visual medium.

Agents proceed to interact with their Environment and perceive the effects and consequences of the Agent Plans they apply as sequences of Situations they recognize and intend to realize. A world state as represented in a Situation can then prompt an Agent Communication Action Event. And the Task Assignment subcategory of Agent Communication may have a multiplicity of goals represented in terms of Situations. In this context, the notion of an Agent's intention to realize a Situation representing either assigned or internally determined goals denotes the Agent's commitment to apply Agent Plans for achieving the goals (see Russel and Norvig [B46] and Tufis and Ganascia [B53]).

(forall (d r) (if (prompts d r)
 (and (Situation d)
 (ERAS-TLO:AgentCommunication r)
)))

(forall (d r) (if (has_as_goal d r)
 (and (TaskAssignment d)
 (Situation r)
)))

(forall (t) (if (TaskAssignment t)
 (>= 1 (#{ s | (and (has_as_goal t s) (Situation s)) }))))

(forall (s ai ac) (if (and (Situation s)
 (Agent ai)
 (recognizes ai s)
 (prompts s ac)
 (or (Explanation ac)
 (Query ac)
 (Answer ac)
 (TaskAssignment ac))))

```
(initiates ai ac)  
))
```

```
(forall ( ta ) (if (TaskAssignment ta)  
  (exists (ar s)  
    (and (Agent ar)  
          (Situation s)  
          (receives ar ta)  
          (has_as_goal ta s)  
          (intends_to_realize ar s))))))
```

```
(forall (d r) (if (is_response_to d r)  
  (and (Answer d)  
        (Query r))))
```

```
(forall (w q ax) (if (and (Answer w)  
  (Query q)  
  (Agent ax)  
  (is_response_to w q)  
  (receives ax q)  
  (initiates ax w))  
  (exists (aq)  
    (and (Agent aq)  
          (initiates aq q)  
          (receives aq w))))))
```

Agents may be members of particular Social Collections, where the Collective binding corresponds to an aggregation of agents grouped together by some common property or social purpose. The Social Collection category is a subcategory of the ERAS-TLO:Collective category and inherits the `is_member_of` relationship between the Entity and Collective concepts.

```
(forall (s) (if (SocialCollection s) (ERAS-TLO:Collective s)))
```

```
(forall (a) (if (Agent a)  
  (exists(s)  
    (and (SocialCollection s)  
          (is_member_of a s))))))
```

An Agent may be a member of one or more Social Collections.

```
(forall (a) (if (Agent a)  
  (>= 1 (#{ s | (and (is_member_of a s) (SocialCollection s)) } ))))
```

Social Collections might influence the way Norms are applied.

```
(forall (d r) (if (influences_norm_applicability d r)  
  (and (SocialCollection d)  
        (Norm r))))
```

A Social Collection may influence the applicability of one or more Norms.

(forall (s) (if (SocialCollection s)
 (>= 1 (# { x | (and (influences_norm_applicability s x) (Norm x)) })))))

A Social Collection may prescribe the context for situations that represent an Agent's perception of its environment.

(forall (d r) (if (prescribes_context_for d r)
 (and (SocialCollection d)
 (Situation r)
)))

The Social Collection category is specialized with the following six subcategories:

(forall (s) (if (Community s) (SocialCollection s)))

(forall (s) (if (Company s) (SocialCollection s)))

(forall (s) (if (Government s) (SocialCollection s)))

(forall (s) (if (Department s) (SocialCollection s)))

(forall (s) (if (Organization s) (SocialCollection s)))

(forall (s) (if (Team s) (SocialCollection s)))

The Social Collection subcategories are defined informally as follows:

- A Community Social Collection is an aggregation of agents grouped together by common properties such as geographic location, ethnic affiliations, or shared values.
- A Company Social Collection is an aggregation of agents as employees of a company.
- A Government Social Collection is an aggregation of agents participating in a governmental system that governs an organized community or state for the purpose of establishing direction, rights, obligations, and control over members of the community or state.
- A Department Social Collection is an aggregation of agents belonging to a subgroup that is part of a larger group, company, or organization.
- An Organization Social Collection is an aggregation of agents belonging to a group of participants with a shared purpose.
- A Team Social Collection is an aggregation of agents formed for some usually short term objective.

Participating team members that are ethically aware autonomous agents need to perceive and account for the types of teams in which they are members. A descriptive characterization of two important team types follows:

(= (Description team_types)
 { self_directed_MAS_agents,
 human_directed })

Agents can be assigned various roles in their Social Collections. In this ontology Agent Role is a subcategory of the ERAS-TLO:Role category and as such characterizes a defined set of connected

behaviors, capabilities, requirements, rights, obligations, and permissions expected of any agent assigned or ascribed the respective agent role. As a subcategory of Role, Agent Role inherits the enacts relationship between Agent and Role but specializes the formalization of the relationship by requiring that an enacted Agent Role be assigned to the Agent.

```
(forall (r) (if (AgentRole r) (ERAS-TLO:Role r)))
```

```
(forall (d r) (if (assigned_to d r)  
  (and (AgentRole d)  
  (Agent r))))
```

An Agent Role may be assigned to zero or more Agents.

```
(forall (r) (if (AgentRole r)  
  (>= 0 (#{ a | (and (assigned_to r a) (Agent a)) }))))
```

```
(forall (d r) (if (enacts d r)  
  (and (AgentRole r)  
  (Agent d)  
  (assigned_to r d))))
```

Agents may enact zero or more Agent Roles.

```
(forall (a) (if (Agent a)  
  (>= 0 (#{ r | (and (enacts a r) (AgentRole r)) }))))
```

Agent Roles may stipulate zero or more Norms that are relevant for an Agent to consider when enacting the role.

```
(forall (d r) (if (stipulates d r)  
  (and (AgentRole d)  
  (Norm r))))
```

```
(forall( r ) (if (AgentRole r)  
  (>= 0 (#{ n | (and (stipulates r n) (Norm n)) }))))
```

4.6 Data Privacy and Protection

Figure 3 shows the UML diagram depicting concepts and relationships for the Data Protection and Privacy (DPP) subdomain. This model focuses on documenting relevant concepts and relationships characterizing the data protection and privacy rules and regulations that shall be observed and upheld by ethical agents and autonomous systems. Some concepts identified in the other subdomains occur here and are identified with an annotated color. An explanation about each concept is expressed in a mix of sentences in natural language and axioms.

A Person is an Agent that is granted a range of specific data subject rights regarding the use and protection of data about themselves. A Person's personal data includes private and public information emanating from life and personal activities such as social and citizenship events or engaging in consumption of services and products.

```
(forall (p) (if (Person p) (and (ERAS-NEP:Agent p) (not (Robot p)))))
```

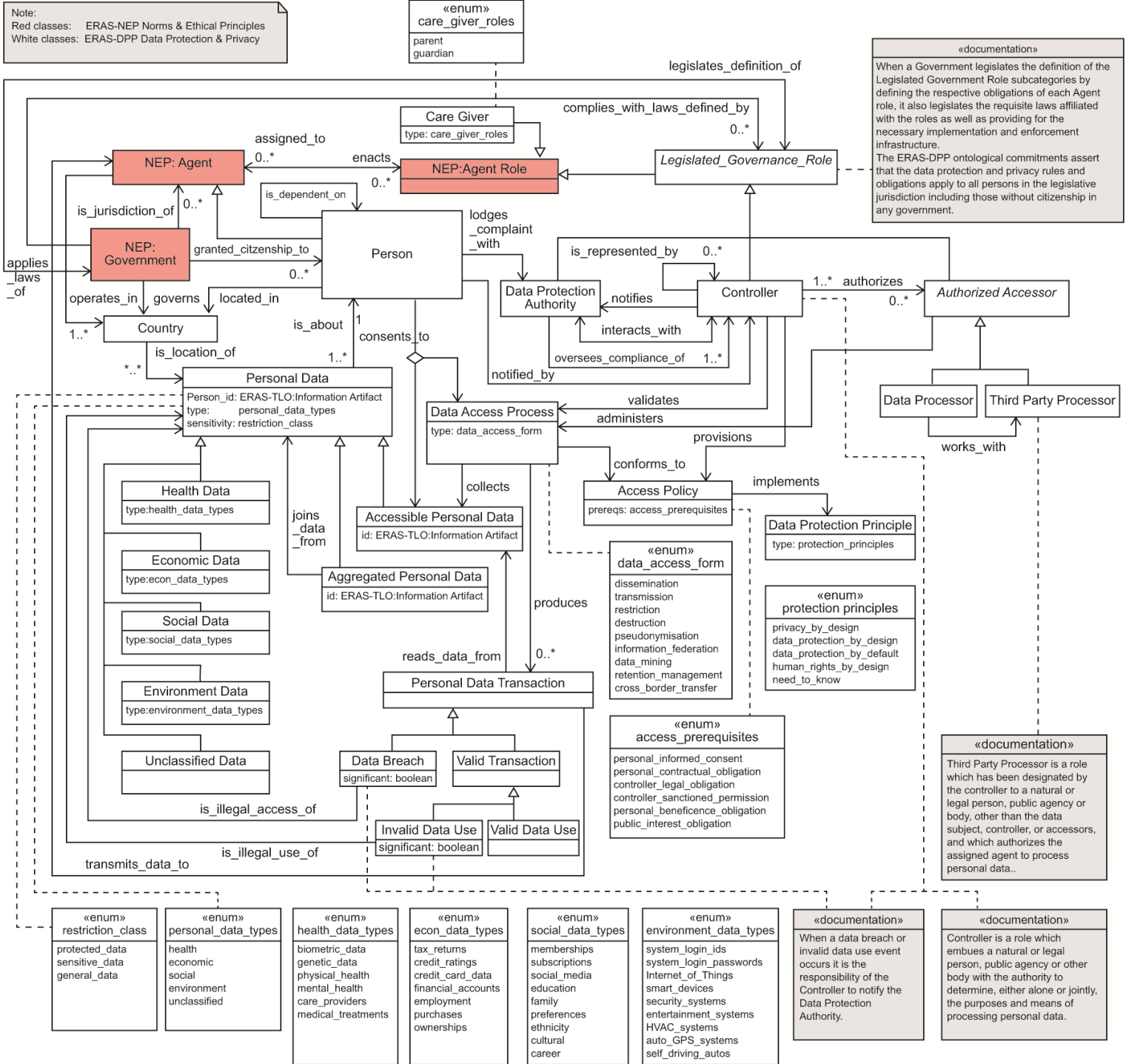


Figure 3 — Data Privacy and Protection UML Diagram

Data privacy is a highly complex and increasingly regulated area of law, in which the regulatory regime is rapidly evolving. No standard can provide unconditional consistency with all applicable laws and regulations, which continue to change rapidly in this area, and may also vary at the local, state and regional level. Users of this standard are responsible for keeping apprised of such laws and regulations.

As a subcategory of the ERAS-NEP:Agent category, the Person concept can be located in physical Continuant categories such as Country. A Person is located in one Country at a specific point in time.

(forall (c) (if (Country c) (ERAS-TLO:Continuant c)))

(forall (d r) (if (located_in d r)
(and (Person d)
(Country r))))

(forall (p) (if (Person p)
(= 1 (#{ c | (and (located_in p c) (Country c) })))))

All people have a set of personal data, and each personal data is about just one person.

(forall (p) (if (Person p)
(exists (d)
(and (PersonalData d)
(is_about d p))))))

(forall (d r) (if (is_about d r)
(and (PersonalData d)
(Person r))))

Personal Data is restricted to be about just one person. The following axiom uses the shorthand notation to express the cardinality restriction:

(forall (d) (if (PersonalData d)
(= 1 (#{ p | (and (is_about d p) (Person p) })))))

The Personal Data about a person may have many distributed and distinct forms so the multiplicity of the is_about relationship to a Person is one to many.

(forall (p) (if (Person p)
(>= 1 (#{ d | (and (is_about d p) (PersonalData d) })))))

Personal Data is any information relating to an identified or identifiable natural person (the data subject) in a personal capacity. The means of identification can be determined, directly or indirectly by name, identification number, location data, or by one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of the data subject. These characteristics are often referred to as Personal Identifying Information.

(forall (d) (if (PersonalData d) (ERAS-TLO:InformationArtifact d)))

The Personal data about a data subject may be distributed and collected across many countries.

(forall (d) (if (PersonalData d)
(exists (c)

```
(and (Country c)
      (is_location_of c d))))
```

```
(forall (d r) (if (is_location_of d r)
                  (and (Country d)
                       (PersonalData r))))
```

```
(forall (d) (if (PersonalData d)
                (>= 1 (#{ c | (and (is_location_of c d) (Country c)) }))))
```

```
(forall (d) (if (Country c)
                (>= 1 (#{ d | (and (is_location_of c d) (PersonalData d)) }))))
```

When realized as an Information Artifact about a Person, the Personal Data will possess a sensitivity attribute that specifies a range of restriction class ratings. A descriptive characterization of such ratings include the following examples:

```
(= (Description restriction_class)
   { protected_data
     sensitive_data
     general_data }))
```

The Personal Data category is specialized with five subcategories as follows:

- *Personal Health Data*: Personal data associated with the health, physiological state and characteristics of the data subject.
(forall (x) (if (HealthData x) (PersonalData x)))
- *Economic Data*: Personal data associated with the economic state and characteristics of the data subject.
(forall (x) (if (EconomicData x) (PersonalData x)))
- *Social Data*: Personal data associated with the sociological state and characteristics of the data subject.
(forall (x) (if (SocialData x) (PersonalData x)))
- *Environment Data*: Personal data associated with information derived from the personal environment inhabited by the data subject.
(forall (x) (if (EnvironmentData x) (PersonalData x)))
- *Unclassified Data*: Miscellaneous Personal data not classified in any of the other subcategories.
(forall (x) (if (UnclassifiedData x) (PersonalData x)))

Each Personal Data subcategory has a data type signifying the range of data classified by the subcategory. Descriptive characterizations of these data type include the following examples:

```
(= (Description health_data_types)
   { biometric_data
     genetic_data
     physical_health
```

```
mental_health  
care_providers  
medical_treatments )))
```

```
( = ( Description econ_data_types)  
  { tax_returns  
    credit_ratings  
    credit_card_data  
    financial_accounts  
    employment  
    purchases  
    ownerships } ) ) )
```

```
( = ( Description social_data_types)  
  { memberships  
    subscriptions  
    social_media  
    education  
    family_preferences  
    ethnicity  
    cultural  
    career } ) ) )
```

```
( = ( Description environment_data_types)  
  { system_login_ids  
    system_login_passwords  
    Internet_of_Things  
    smart_devices  
    security_systems  
    entertainment_systems  
    heating_ventilation_and_air_conditioning_systems [HVAC8]  
    auto_GPS_systems  
    self_driving_autos } ) ) )
```

Aggregated Personal Data is also Personal data that has been collected, compiled, or data mined across multiple sources including public and private databases, social media, web sites and personal artifacts that can be used to infer and reveal new or previously unpublished and unavailable personal information about a data subject.

```
(forall (x) (if (AggregatedPersonalData x) (PersonalData x)))
```

Aggregated Personal Data may or may not exist for any particular person and there may be multiple cases of Aggregated Personal Data for a person.

```
(forall (a) (if (Person a)  
  ( >= 0 ( #{ g | (and (is_about g a) (AggregatedPersonalData g)) } ) ) ) ) )
```

⁸ The acronym HVAC is an abbreviation for heating, ventilation, and air conditioning.

The Aggregated Personal Data category characterizes the logical joining of data among and across the data subcategories that comprise the personal data and information about individuals.

```
(forall (d r) (if (joins_data_from d r)
  (and (AggregatedPersonalData d)
    (PersonalData r))))
```

```
(forall (a) (if (AggregatedPersonalData a)
  (exists (d)
    (and (PersonalData d)
      (joins_data_from a d))))))
```

Persons may grant access to their Accessible Personal Data for specific data transactions and valid data usages. Accessible Personal Data is the subcategory of Personal Data which classifies that portion of data for which the data subject may grant consent for such access.

```
(forall (d) (if (AccessiblePersonalData d) (PersonalData d)))
```

A Person may grant consent to access their Accessible Personal Data if the Person is not dependent on a Person assigned the Care Giver role of a parent or a guardian.

A Person may be assigned and enact the role of Care Giver as a Parent of or as a Guardian of a Person.

```
(forall (g) (if (CareGiver g) (ERAS-NEP:AgentRole g)))
```

A descriptive characterization of the types of Care Giver roles includes the following enumerated examples.

```
(= ( Description care_giver_roles)
  { parent
    guardian } )
```

A dependency relationship is established between the person enacting the Care Giver role and the person for whom the Care Giver is a parent or guardian.

```
(forall (d r) (if (is_dependent_on d r)
  (and (Person d)
    (Person r)
    (not (= d r))))))
```

A Person may be dependent upon another Person who is enacting an Agent Role of Care Giver for the Person.

```
(forall (p g) (if (and (Person p)
  (Person g)
  (not (= g p))
  (is_dependent_on p g))
  (exists (c)
    (and (AgentRole c)
      (CareGiver c)
      (enacts g c))))))
```

Consent to access a Person's Accessible Personal Data involves permitting a Data Access Process to collect the data intended for the authorized usage. A Data Access Process is a sequence of operations that have been authorized and validated by Agents enacting the relevant Legal Governance Roles in effect for the Person's circumstances. A validated Data Access Process may then generate a Personal Data Transaction to access the Accessible Personal Data of the Person that consented to the transaction.

(forall (p) (if (DataAccessProcess p) (ERAS-TLO:Process p)))

(forall (x) (if (DataAccessProcess x)
 (exists (a)
 (and (AccessiblePersonalData a)
 (collects x a))))))

(forall (d r) (if (collects d r)
 (and (DataAccessProcess d)
 (AccessiblePersonalData r))))

(forall (p ad dp) (if (consents_to p dp ad)
 (and (Person p)
 (DataAccessProcess dp)
 (AccessiblePersonalData ad)
 (collects dp ad)
 (is_about ad p)
)))

A Person who is not formally dependent upon another Person may consent to the access of their Accessible Personal Data by a valid Data Access Process.

(forall (p pd ap) (if (and (Person p)
 (AccessiblePersonalData pd)
 (DataAccessProcess ap)
 (collects ap pd)
 (is_about pd p)
 (consents_to p pd ap))
 (not (exists (g)
 (and (Person g)
 (not (= g p))
 (is_dependent_on p g))))))

If a Person is a minor or the ward of a guardian, the parent or guardian is the person who may grant access to the Accessible Personal Data of the minor or ward.

(forall (p g c pd ap) (if (and (Person p)
 (Person g)
 (not (= p g))
 (AgentRole c)
 (CareGiver c)
 (enacts g c)
 (AccessiblePersonalData pd)
 (is_about pd p)

```
(DataAccessProcess ap)  
(collects ap pd)  
(is_dependent_on p g)  
(consents_to g pd ap)))
```

A Data Access Process may exhibit many forms of access. A descriptive characterization of such forms includes the following examples:

```
( = ( Description data_access_form)  
  { dissemination  
    transmission  
    restriction  
    destruction  
    pseudonymisation  
    information_federation  
    data_mining  
    retention_management  
    cross_border_transfer  
  } )
```

A Data Access Process may produce many Personal Data Transactions.

The Personal Data Transaction category classifies data transactions initiated by a data access process to access and operate on the Accessible Personal Data of a data subject.

```
(forall (p) (if (PersonalDataTransaction p) (ERAS-TLO:Process p)))
```

```
(forall (d r) (if (produces d r)  
  (and (DataAccessProcess d)  
    (PersonalDataTransaction r)  
  )))
```

```
(forall (d) (if (DataAccessProcess d)  
  ( >= 0 (# { r | (and (produces d r) (PersonalDataTransaction r)) } ) )))
```

```
(forall (d r) (if (reads_data_from d r)  
  (and (PersonalDataTransaction d)  
    (AccessiblePersonalData r)  
  )))
```

A Personal Data Transaction that is produced by a Data Access Process transmits data to some Agent.

```
(forall (d r) (if (transmits_data_to d r)  
  (and (PersonalDataTransaction d)  
    (ERAS-NEP:Agent r)  
  )))
```

The Data Privacy and Protection ERAS subdomain commits to a number of Agent Roles defined for the purpose of specifying responsibilities, permissions, and obligations associated with the duties of securing the protection and privacy of personal data. In many geographical and national regions these roles will be

legislated by the regional government agencies. Users of the standard are responsible for verifying whether the Agent Roles are regulated by the relevant local or regional government, and how that impacts the responsibilities, permissions, and obligations, consistent with all applicable laws and regulations.

(forall (g) (if (LegislatedGovernanceRole g) (ERAS-NEP:AgentRole g)))

The *Legislated Governance Role* category is an abstract conceptualization⁹ that is specialized with three subcategories, Data Protection Authority, Controller, and Authorized Accessor.

(forall (r) (if (DataProtectionAuthority r) (LegislatedGovernanceRole r)))

(forall (r) (if (Controller r) (LegislatedGovernanceRole r)))

(forall (r) (if (AuthorizedAccessor r) (LegislatedGovernanceRole r)))

The specialized roles have the following responsibilities, authorities, and obligations:

- *Data Protection Authority (DPA)*: The principal supervisory authority responsible for consistent application and enforcement of personal data and privacy protection policies and directives. A DPA becomes the main point of contact for participating stakeholder communities.
- *Controller*: A natural or legal person, public agency or other body with the authority to determine, either alone or jointly, the purposes and means of processing personal data.
- *Authorized Accessor*: A generic or abstract agent role subclass representing common properties and relationships assigned to persons, natural or legal, public authorities or agencies other than data subjects, and controllers that have been authorized by a controller to process personal data. Subclasses of this abstract concept represent the specific Accessor roles that may be assigned to agents.

A Person suspecting that their Personal Data may have been inappropriately accessed or used may lodge a complaint with their associated Data Protection Authority.

(forall(d r) (if (lodges_complaint_with d r)
(and (Person d)
(DataProtectionAuthority r))))

(forall(p r) (if (and (Person p)
(DataProtectionAuthority r)
(lodges_complaint_with p r))
(exists(a)
(and (ERAS-NEP:Agent a)
(assigned_to r a)
(enacts a r))))))

The *Authorized Accessor* abstract role has two specializations, Data Processor, and Third Party Data Processor.

(forall (r) (if (DataProcessor r) (AuthorizedAccessor r)))

(forall (r) (if (ThirdPartyProcessor r) (AuthorizedAccessor r)))

⁹ Following UML standard semantics, an abstract UML concept is denoted in Italics and is intended to designate a M1 model classification that does not have M0 model instances.

The two specialized subcategories of *Authorized Accessor* have the following capabilities, responsibilities, and obligations.

- *Data Processor*: A kind of Authorized Accessor that is a natural or legal person, public authority, agency, or other body that processes personal data as authorized by the Controller.
- *Third Party Processor*: A kind of Authorized Accessor that is natural or legal person, public authority, or body other than the data subject, controller, or accessors that have been authorized by a Controller to process personal data on behalf of the Controller while working with a Data Processor with whom it shares personal data.

An Agent enacting the role of Data Protection Authority oversees the compliance of one or more Controller Agents.

(forall (d) (if (DataProtectionAuthority d)
(≥ 1 (# { c | (and (oversees_compliance_of d c) (Controller c)) }))))

(forall (d) (if (DataProtectionAuthority d)
(exists (c)
(and (Controller c)
(oversees_compliance_of d c))))))

(forall (d r) (if (oversees_compliance_of d r)
(and (DataProtectionAuthority d)
(Controller r))))

A Controller Agent is responsible for interacting with and notifying the Data Protection Authority when data is accessed illegally through a data breach. The Person whose Personal Data was involved with a data breach is also notified by the Controller.

(forall (d r) (if (interacts_with d r)
(or (and (Controller r)
(DataProtectionAuthority d))
(and (Controller d)
(DataProtectionAuthority r))))))

(forall (d r) (if (notifies d r)
(and (Controller d)
(DataProtectionAuthority r))))

(forall (d r) (if (notified_by d r)
(and (Person d)
(Controller r))))

A Controller is permitted to delegate an Authorized Accessor role to other Agents as necessary.

(forall (d r) (if (authorizes d r)
(and (Controller d)
(AuthorizedAccessor r))))

(forall (c) (if (Controller c)
(≥ 0 (# { a | (and (authorizes c a) (AuthorizedAccessor a)) }))))

An Agent enacting one of the Authorized Accessor subcategory roles was authorized by at least one Controller.

```
(forall (a) (if (AuthorizedAccessor a)
  ( >= 1 (#{ c | (and (authorizes c a) (Controller c)) } ))))
```

```
(forall (a) (if (AuthorizedAccessor a)
  (exists (c)
    (and (Controller c)
      (authorizes c a)
    )))
```

A Controller may designate other Controllers to act as a representative of the designating Controller regarding matters of data privacy and protection obligations assigned to the designating Controller.

```
(forall (x y) (if (is_represented_by x y)
  (and (Controller x)
    (Controller y)
    (not (= x y))))
```

```
(forall (c) (if (Controller c)
  ( >= 0 (#{ a | (and (is_represented_by c a) (Controller a)) } ))))
```

Agents enacting the role of Controller validate the Data Access Processes used to collect Accessible Personal Data.

```
(forall (d r) (if (validates d r)
  (and (Controller d)
    (DataAccessProcess r))
  ))
```

```
(forall (r) (if (DataAccessProcess r)
  (exists (c)
    (and (Controller c)
      (validates c r))))
```

Agents enacting either of the two subclass roles of *Authorized Accessor*, that is as a Data Processor or a Third Party Processor, administer the Data Access Processes that have been validated by the Controller that authorized their Access. Note that the Authorized Accessor abstract role cannot be directly assigned to an Agent, only the subcategory roles can be assigned.

```
(forall (d r) (if (administers d r)
  (and (AuthorizedAccessor d)
    (or (DataProcessor d)
      (ThirdPartyProcessor d))
    (DataAccessProcess r)
    (exists (c)
      (and (Controller c)
        (validates c r)
        (authorizes c d))))))
```

Data Processors may work with Third Party Processors with whom they share Personal Data.

```
(forall (d r) (if (works_with d r)
  (and (DataProcessor d)
    (ThirdPartyProcessor r))))
```

```
(forall (r) (if (ThirdPartyProcessor r)
  (exists (d)
    (and (DataProcessor d)
      (works_with d r))))))
```

Each Data Access Process conforms to an Access Policy that implements a Data Protection Principle. An Access Policy is a data privacy and protection policy that specifies the requirements and prerequisites necessary to control and protect the collection, access, and use of personal data about the data subject.

```
(forall (x) (if (AccessPolicy x) (ERAS-TLO:Method x)))
```

Agents enacting the role of Controller provision their data protection and privacy control infrastructure with required Access Policies.

```
(forall (d r) (if (provisions d r)
  (and (Controller d)
    (AccessPolicy r))))
```

```
(forall (a) (if (AccessPolicy a)
  (exists (c)
    (and (Controller c)
      (provisions c a))))))
```

```
(forall (d r) (if (conforms_to d r)
  (and (DataAccessProcess d)
    (AccessPolicy r))))
```

```
(forall (d) (if (DataAccessProcess d)
  (exists (a)
    (and (AccessPolicy a)
      (conforms_to d a))))))
```

A descriptive characterization of access prerequisites to be enforced by an Access Policy includes the following enumerated examples:

```
(= (Description access_prerequisites)
  { personal_informed_consent
    personal_contractual_obligation
    controller_legal_obligation
    controller_sanctioned_permission
    personal_beneficence_obligation
    public_interest_obligation
  } )
```

A Data Protection Principle articulates general guidelines intended to enable the protection and use of personal data across evolving technology and multiple stakeholder communities. Such principles are implemented by Access Policies, which are in consonance with validated Data Access Processes, when collecting Accessible Personal Data.

```
(forall (x) (if (DataProtectionPrinciple x) (ERAS-TLO:Method x)))
```

```
(forall (d r) (if (implements d r)  
  (and (AccessPolicy d)  
    (DataProtectionPrinciple r ))))
```

```
(forall (r) (if (AccessPolicy r)  
  (exists (p)  
    (and (DataProtectionPrinciple p)  
      (implements r p)  
    )))
```

A descriptive characterization of protection principles, to be implemented, includes the following enumerated examples:

```
(= ( Description protection_principles)  
  { privacy_by_design  
    data_protection_by_design  
    data_protection_by_default  
    human_rights_by_design  
    need_to_know10  
  } )
```

The Personal Data Transaction category is specialized with two subcategories: Valid Transaction and Data Breach.

```
(forall (p) (if (ValidTransaction p) (PersonalDataTransaction p)))
```

```
(forall (p) (if (DataBreach p) (PersonalDataTransaction p)))
```

The Valid Transaction category classifies¹¹ a data transaction initiated by a data access process that accesses and operates on the Accessible Personal Data of the data subject with the consent of the data subject and under the auspices of an Authorized Accessor agent.

```
(forall (v) (if (ValidTransaction v)  
  (exists (p ad dp aa)  
    (and (Person p)  
      (AccessiblePersonalData ad)  
      (DataAccessProcess dp)  
      (AuthorizedAccessor aa)  
      (is_about ad p)
```

¹⁰ The need to know data protection principle refers to a policy normally applied within confidential information contexts, such as HR human resource departments or research and development laboratories of enterprises, where personal information such as salaries and project research notes are only released to others based on their need to know as established by the roles they enact.

¹¹ The verb “classifies” is used here to express the UML concept of classification where the UML class classifies the concept in terms of its attributes and relationships.

```
(consents_to p dp ad)  
(administers aa dp)  
(produces dp v)  
(reads_data_from v ad)  
))))
```

The Valid Transaction category is specialized with two subcategories: Valid Data Use and Invalid Data Use.

```
(forall (p) (if (ValidDataUse p) (ValidTransaction p)))
```

```
(forall (p) (if (InvalidDataUse p) (ValidTransaction p)))
```

The Valid Data Use category classifies a Personal Data Transaction process accessing Accessible Personal Data in which the process used, the data accessed, and the use of that data satisfies all data privacy and protection constraints in place for the Person the data is about. This means that the Valid Data Use subclass of a Valid Transaction presents a legal access and legal use of the Accessible Personal Data that is accessed.

```
(forall (vu) (if (ValidDataUse vu)  
  (exists (p ad a)  
    (and (PersonalDataTransaction vu)  
         (ValidTransaction vu )  
         (Person p)  
         (AccessiblePersonalData ad)  
         (is_about ad p)  
         (reads_data_from vu ad)  
         (not (InvalidDataUse vu))  
         (not (DataBreach vu))  
         (ERAS-NEP:Agent a)  
         (transmits_data_to vu a)  
      )))
```

The Invalid Data Use category classifies a data transaction accessing Accessible Personal Data that satisfies the legal access requirements of the data access but permits an illegal use of the data as prescribed by the data privacy and protection constraints in effect.

```
(forall (d r) (if (is_illegal_use_of d r)  
  (and (InvalidDataUse d)  
       (PersonalData r))))
```

The following axiom duplicates some of the more specific properties in the axiom that follows it.

```
(forall (x) (if (InvalidDataUse x)  
  (exists (p)  
    (and (PersonalData p)  
         (is_illegal_use_of x p))))))
```

```
(forall (iu) (if (InvalidDataUse iu)  
  (exists (p da c ad)  
    (and (Person p)
```

```
(AccessiblePersonalData ad)  
(is_about ad p)  
(reads_data_from iu ad)  
(not (DataBreach iu))  
(is_illegal_use_of iu ad)  
(Controller c)  
(DataProtectionAuthority da)  
(notifies c da)  
(notified_by p c))))
```

The Data Breach category classifies an incident in which sensitive, protected, or confidential Personal Data is copied, transmitted, viewed, stolen, or used by an agent, or a personal data transaction on behalf of an agent, that is unauthorized to do so and for which the data subject has not granted consent to access. Such an incident constitutes an illegal access of the Personal Data and requires notification of the data subject and the Data Protection Authority by the responsible Controller.

Each local and regional governmental jurisdiction may have its own definition for what qualifies as a Data Breach, with differing definitions regarding what type of information is sensitive, protected or confidential Personal Data. Users of this standard should consult legal counsel to determine whether an illegal or improper Data Breach has occurred, and how to respond based on the applicable laws and regulations of the jurisdiction.

```
(forall (d) (if (is_illegal_access_of d r)  
  (and (DataBreach d)  
    (PersonalData r)  
  )))
```

The following axiom duplicates some of the more specific properties in the axiom that follows it.

```
(forall (d) (if (DataBreach d)  
  (exists (p)  
    (and (PersonalData p)  
      (is_illegal_access_of d p))))))
```

```
(forall (db) (if (DataBreach db)  
  (exists (p ad dp c pa)  
    (and (Person p)  
      (AccessiblePersonalData ad)  
      (DataAccessProcess dp)  
      (is_about ad p)  
      (is_illegal_access_of db ad)  
      (Controller c)  
      (DataProtectionAuthority pa)  
      (or (not (consents_to p dp ad))  
        (not (exists aa)  
          (and (AuthorizedAccessor aa)  
            (administers aa dp))))))  
    (notifies c pa)  
    (notified_by p c)  
    (interacts_with c pa)  
    (interacts_with pa c))))))
```

As a subcategory of Agent Role, the Legislated Governance Role abstract category inherits the “assigned_to” and “enacts” relationships between Agent and Agent Role. Consequently, the four subcategories of the Legislated Governance Role abstract category may be assigned to Agents and assigned Agents may enact one or more of the roles.

Agents enacting these roles can operate in one or more countries as they fulfill the responsibilities and obligations defined for the roles.

```
(forall (d r) (if (operates_in d r)
  (and (ERAS-NEP:Agent d)
    (Country r)
    (exists (g)
      (and (LegislatedGovernanceRole g)
        (enacts d g))))))
```

```
(forall (a) (if (ERAS-NEP:Agent a)
  (>= 1 (# { c | (and(operates_in a c) (Country c) } ))))
```

Countries are governed by Government Social Collections and Governments may legislate the definition of the set of Legislated Governance Roles. When a Government legislates the definition of the Legislated Governance Role subcategories by defining the respective obligations of each Agent role, it also legislates the requisite laws affiliated with the roles as well as providing the necessary implementation and enforcement infrastructure.

```
(forall (d r) (if (governs d r)
  (and (ERAS-NEP:Government d)
    (Country r)
  )))
```

```
(forall (d r) (if (legislates_definition_of d r)
  (and (ERAS-NEP:Government d)
    (LegislatedGovernanceRole r))))
```

```
(forall (d r) (if (applies_laws_of d r)
  (and (LegislatedGovernanceRole d)
    (ERAS-NEP:Government r))))
```

```
(forall (g r) (if (and(ERAS-NEP:Government g)
  (LegislatedGovernanceRole r)
  (legislates_definition_of g r))
  (applies_laws_of r g)))
```

```
(forall (r g) (if (and (ERAS-NEP:Government g)
  (LegislatedGovernanceRole r)
  (applies_laws_of r g))
  (legislates_definition_of g r)))
```

```
(forall (g c dp) (if (and (ERAS-NEP:Government g)
  (Controller c)
  (LegislatedGovernanceRole c)
```



```
(DataAccessProcess dp)
(legislates_definition_of g c)
(validates c dp))
(exists (ap dpp)
  (and (AccessPolicy ap)
    (DataProtectionPrinciple dpp)
    (provisions c ap)
    (conforms_to dp ap)
    (implements ap dpp))
  )))
```

A Government may choose to adopt and comply with the laws and obligations defined for the Legislated Governance Roles that were defined and established by another Government.

```
(forall (d r) (if (complies_with_laws_defined_by d r)
  (and (ERAS-NEP:Government d)
    (LegislatedGovernanceRole r)
    (not (legislates_definition_of d r))
  )))
```

A Government may establish the legal jurisdiction of zero or more Agents and it may grant citizenship to zero or more Persons.

In the context of the DPP subdomain, a Person's citizenship relation merely establishes possible Personal Data derived from one's citizenship status and deemed relevant for protection by rules prescribed within the Legislated Governance Role obligations.

```
(forall (d r) (if (is_jurisdiction_of d r)
  (and (ERAS-NEP:Government d)
    (ERAS-NEP:Agent r))))
```

```
(forall (g) (if (ERAS-NEP:Government g)
  (>= 0 (#{ a | (and (is_jurisdiction_of g a)(ERAS-NEP:Agent a)) } ))))
```

```
(forall (d r) (if (granted_citizenship_to d r)
  (and (ERAS-NEP:Government d)
    (Person r))))
```

```
(forall (g) (if (ERAS-NEP:Government g)
  (>= 0 (#{ p | (and(granted_citizenship_to g p) (Person p)) } ))))
```

DPP laws are typically applied by the Government that governs the Country in which the Personal Data is located without regard to citizenship, applying to citizens as well as those who are stateless or without citizenship in any Government.

```
(forall (p pd) (if (and (Person p)
  (AccessiblePersonalData pd)
  (is_about pd p)
  (not (exists (g)
    (and (ERAS-NEP:Government g)
      (granted_citizenship_to g p))))
  )))
```

```
(forall (dap ap dpp)
  (if (and (DataAccessProcess dap)
    (AccessPolicy ap)
    (DataProtectionPrinciple dpp)
    (conforms_to dap ap)
    (implements ap dpp)
    (collects dap pd )
    (consents_to p dap pd))
    (exists (ai aj ak dpa rc raa)
      (and (ERAS-NEP:Agent ai)
        (ERAS-NEP:Agent aj)
        (ERAS-NEP:Agent ak)
        (DataProtectionAuthority dpa)
        (enacts ai dpa)
        (Controller rc)
        (enacts aj rc)
        (AuthorizedAccessor raa)
        (enacts ak raa)
        (validates rc dap)
        (provisions rc ap)
        (administers raa dap))))))
```

4.7 Transparency and Accountability

Figure 4 shows the UML diagram that captures the main concepts and relationships identified in the Transparency and Accountability (TA) subdomain. This model intends to capture the concepts and relationships necessary to enable ethical autonomous systems with capabilities that provide informative explanations for plans and associated actions. Some concepts identified in the other subdomains are identified and shown in a specific color. An explanation about each concept is expressed in a mix of sentences in natural language and axioms.

Ethically-aware Agents need to have the ability to be transparent in their interactions with other agents. An agent qualifies as an autonomous transparent agent if it is enabled with an always available mechanism capable of reporting its behavior, intentions, perceptions, goals, and constraints in a manner that permits authorized users and collaborating agents to understand its past and expected future behavior.

As specified in the NEP subdomain, autonomous agents can interact with other agents by initiating and receiving Agent Communication Action Events to transmit and exchange Information Artifacts. One of the subcategories of Agent Communication Action Events is the Explanation subcategory. The Transparency and Accountability (TA) subdomain defines additional relationships for the Explanation concept.

Agents may receive many requests for Explanations and they are accountable for the Explanations they provide in response.

```
(forall (d r) (if (is_accountable_for d r)
  (and (ERAS-NEP:Agent d)
    (ERAS-NEP:Explanation r))))

(forall (a) (if (ERAS-NEP:Agent a)
  (>= 0 (#{ e | (and (is_accountable_for a e) (ERAS-NEP:Explanation e)) } ))))
```

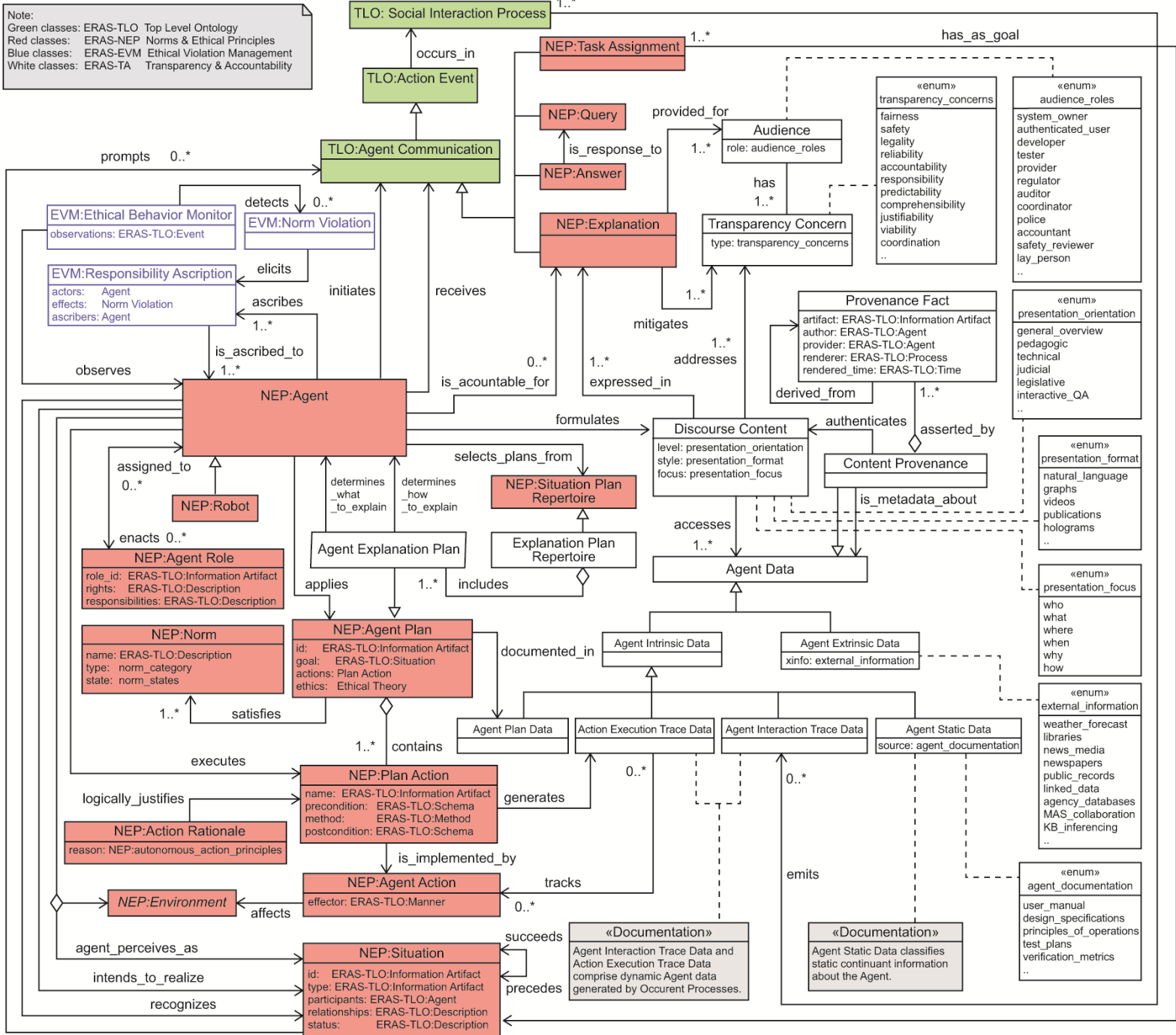


Figure 4 — Transparency and Accountability UML Diagram

Any agent explanation is based on an explanation plan repertoire that contains a collection of action plan templates that characterizes a set of principles to guide agent plan and action selection for responding to requests for explanations about agent behaviors and capabilities. The Agent Explanation plans included in the Explanation Plan Repertoire are specifications, partial or complete, of agent action sequences that determine what and how to formulate explanations regarding agent capabilities, and past or future behaviors. The Explanation Plan Repertoire is a subcategory of the Situation Plan Repertoire concept.

(forall (x) (if (ExplanationPlanRepertoire x) (ERAS-NEP:SituationPlanRepertoire x)))

As a subclass of the Situation Plan Repertoire, the Explanation Plan Repertoire inherits the `selects_plans_from` relationship between an Agent and the Explanation Plan Repertoire. The Explanation Plan Repertoire includes a collection of Agent Explanation Plans.

(forall (x) (if (AgentExplanationPlan x) (ERAS-NEP:AgentPlan x)))

(forall (d r) (if (includes d r)
 (and (ExplanationPlanRepertoire d)
 (AgentExplanationPlan r))))

(forall (er) (if (ExplanationPlanRepertoire er)
 (>= 1 (#{ ep | (and (includes er ep) (AgentExplanationPlan ep)) }))))

(forall (d) (if (ExplanationPlanRepertoire d)
 (exists(p)
 (and (AgentExplanationPlan p)
 (includes d p))))))

Agent Explanation Plans determine what and how to formulate Explanations requested by Agents.

(forall (d r) (if (determines_what_to_explain d r)
 (and (AgentExplanationPlan d)
 (ERAS-NEP:Agent r))))

(forall (d r) (if (determines_how_to_explain d r)
 (and (AgentExplanationPlan d)
 (ERAS-NEP:Agent r))))

(forall (x) (if (AgentExplanationPlan x)
 (exists (a)
 (and (ERAS-NEP:Agent a)
 (determines_how_to_explain x a)
 (determines_what_to_explain x a))))))

Responses to requests for Agent Explanations will need to address the Transparency Concerns of the Audience involved with the Explanation. Transparency Concerns are Explanation topics and themes that underlie the reasons that motivate requests for explanations of Agent behaviors.

Transparency Concerns are subcategories of the ERAS-TLO Property category.

(forall (x) (if (TransparencyConcern x) (ERAS-TLO:Property x)))

A descriptive characterization of Transparency Concerns that motivate Agent formulation of Explanations intended for the target Audience includes the following enumerated examples:

```
( = ( Description transparency_concerns)
  { fairness
    safety
    legality
    reliability
    accountability
    responsibility
    predictability
    comprehensibility
    justifiability
    viability
    coordination
  } )
```

Agent Explanations may mitigate one or more Transparency Concerns.

```
(forall (d) (if (ERAS-NEP:Explanation d)
  (exists(r)
    (and (TransparencyConcern r)
      (mitigates d r))))))
```

```
(forall (d r) (if (mitigates d r)
  (and (ERAS-NEP:Explanation d)
    (TransparencyConcern r))))
```

```
(forall (e) (if (ERAS-NEP:Explanation e)
  (>= 1 (#{ c | (and (mitigates e c) (TransparencyConcern c) } ))))
```

Explanations formulated by an Agent are provided for an Audience of one or more Agents that are enacting shared Audience Roles. Audience Agents have one or more Transparency Concerns. However, since an Audience's collective membership of Agents all share the same Audience Role, they will have compatible Transparency Concerns that can be mitigated by an Explanation. Since Agents may be assigned more than one role, it is feasible for an Agent to be a member of more than one Audience seeking an Explanation from an autonomous system.

```
(forall (x) (if (Audience x) (ERAS-NEP:SocialCollection x)))
```

```
(forall (x) (if (Audience x)
  (exists(r)
    (and (TransparencyConcern r)
      (has x r))))))
```

```
(forall (d r) (if (has d r)
  (and (Audience d)
    (TransparencyConcern r))))
```

```
(forall(a) (if (Audience a)
```

(>= 1 (# { c | (and (has a c) (TransparencyConcern c) })))

(forall (d) (if (ERAS-NEP:Explanation d)
(exists(r)
(and (Audience r)
(provided_for d r))))))

(forall (d r) (if (provided_for d r)
(and (ERAS-NEP:Explanation d)
(Audience r))))

(forall(e) (if(ERAS-NEP:Explanation e)
(>= 1 (# { a | (and (provided_for e a) (Audience a) })))

(forall (a r) (if (and (Audience a)
(ERAS-NEP:Agent r)
(is_member_of r a)
(and (assigned_to (role a) r)
(enacts r (role a))))))

A descriptive characterization of Audience Roles enacted by Agents requesting Explanations of an Agent's behavior include the following enumerated examples:

(= (Description audience_roles)
{ system_owner
authenticated_user
lay_person
developer
tester
provider
regulator
auditor
coordinator
police
accountant
safety_reviewer})

An Agent formulates Discourse Content to use in its communication back to the Explanation requesters. The Discourse Content is formulated by the Agent to address particular Transparency Concerns and is expressed in Explanations to particular Audiences.

The Discourse Content category is a subcategory of the ERAS-TLO Information Artifact category.

(forall (x) (if (DiscourseContent x) (ERAS-TLO:InformationArtifact x)))

(forall (x) (if (DiscourseContent x)
(exists(d)
(and (ERAS-NEP:Agent d)
(formulates d x))))))

An Explanation's Discourse Content addresses one or more Transparency Concerns.

```
(forall (d) (if (DiscourseContent d)
  (exists(r)
    (and (TransparencyConcern r)
      (addresses d r))))))

(forall (d r) (if (addresses d r)
  (and (DiscourseContent d)
    (TransparencyConcern r))))

(forall (d) (if (DiscourseContent d)
  (>= 1 (#{ c | (and (addresses d c) (TransparencyConcern c)) } ))))
```

And the Discourse Content is expressed in one or more Explanations.

```
(forall (d) (if (DiscourseContent d)
  (exists(r)
    (and (ERAS-NEP:Explanation r)
      (expressed_in d r))))))

(forall (d r) (if (expressed_in d r)
  (and (DiscourseContent d)
    (ERAS-NEP:Explanation r))))

(forall (d) (if (DiscourseContent d)
  (>= 1 (#{ e | (and (expressed_in d e)(ERAS-NEP:Explanation e)) } ))))

(forall (x) (if (DiscourseContent x)
  (exists (a e c d)
    (and (ERAS-NEP:Agent a)
      (ERAS-NEP:Explanation e)
      (TransparencyConcern c)
      (Audience d)
      (formulates a x)
      (expressed_in x e)
      (addresses x c)
      (has d c)
      (provided_for e d)
      (mitigates e c))))))

(forall (e ae s) (if (and (ERAS-NEP:Explanation e)
  (ERAS-NEP:Agent ae)
  (Situation s)
  (prompts s e)
  (receives ae e)
  (is_accountable_for ae e))
  (exists (xr ar dc)
    (and (ERAS-NEP:Agent xr)
      (not (= xr ae))
```

(Audience ar)
(is_member_of xr ar)
(DiscourseContent dc)
(formulates ae dc)
(expressed_in dc e)
(provided_for e ar)
(initiates ae e)
(receives xr e))))))

When formulating a Discourse Content to be expressed in an Explanation, the Agent takes into account the Audience Roles of the Audience agents and their respective Transparency Concerns to frame the Discourse Content in terms of three property dimensions: presentation orientation, presentation form, and presentation focus.

The presentation orientation property aligns the level and type of data in the information content for the discourse Explanation and enumerates the levels of technical specificity used to guide the composition and delivery of requested explanations. A descriptive characterization of presentation orientation levels includes the following enumerated examples:

```
( = ( Description presentation_orientation )  
  { general_overview  
    pedagogic  
    technical  
    judicial  
    legislative  
    interactive_QA  
  } )
```

The presentation format property determines the most appropriate format in which to present the Discourse Content. A descriptive characterization of feasible presentation formats includes the following enumerated examples:

```
( = ( Description presentation_format )  
  { natural_language  
    graphs  
    videos  
    publications  
    holograms  
  } )
```

The presentation focus property determines the illocutionary directives that frame the focus of the explanation request for the Discourse Content response. A descriptive characterization of presentation focus includes the following enumerated examples:

```
( = ( Description presentation_focus )  
  { who  
    what  
    where  
    when  
    why  
    how  
  } )
```


When formulating the Discourse Content for an Explanation, Autonomous Agents have access to a variety of information sources in which to explain and account for their past or future behaviors. The Agent Data category classifies the generic sources of these data.

The Agent Data category has two data subcategories and one metadata subcategory as sources of information for Discourse Content formulation. The subcategories are Agent Intrinsic Data, Agent Extrinsic Data, and the Content Provenance metadata category.

(forall (x) (if (AgentData x) (ERAS-TLO:InformationArtifact x)))

(forall (x) (if (AgentIntrinsicData x) (AgentData x)))

(forall (x) (if (AgentExtrinsicData x) (AgentData x)))

(forall (x) (if (ContentProvenance x) (AgentData x)))

As an Agent formulates the Discourse Content to be expressed in a requested Explanation, one or more of these data sources may be accessed.

(forall (d r) (if (accesses d r)
 (and (DiscourseContent d)
 (AgentData r))))

(forall (d) (if (DiscourseContent d)
 (>= 1 (#{ r | (and (accesses d r) (AgentData r) })))))

The Agent Intrinsic Data category classifies data that is generated by or composed for an Agent and represents information that is self-referencing and about the Agent. The Agent Intrinsic concept has four subcategories: Agent Plan Data, Action Execution Trace Data, Agent Interaction Trace Data, and Agent Static Data.

The Agent Extrinsic Data category classifies data that is not directly about or affiliated with an Agent but which is about external world circumstances in the environment in which the Agent is situated.

The Content Provenance metadata provides information about the Agents and Processes involved in the generation and composition of the respective sources of Agent Data formulated as Discourse Content. This information can be used to assess the quality, reliability, and trustworthiness of the subject data provided in an Agent's Explanation.

The Agent Plan Data category is a subcategory of Agent Intrinsic Data. It classifies the information that documents the methods and procedures of Agent Plans and provides access to Agent Plan Data for the formalization of an Explanation of Agent behavior.

(forall (x) (if (AgentPlanData x) (AgentIntrinsicData x)))

(forall(d r) (if (documented_in d r)
 (and (ERAS-NEP:AgentPlan d)
 (AgentPlanData r))))

When an agent is interacting with others and while acting within its situated environment, it produces internal and external information describing the dynamic and episodic sequence of agent actions, interactions and behaviors. This includes agent perceptions of situational environments, and agent plans for achieving selected objectives prompted by those perceptions.

As Agents apply Agent Plans and execute Plan Actions, internal information in the form of an Action Execution Trace is generated. The Action Execution Trace Data category classifies this episodic sequence history of the prerequisites and consequences of Plan Actions applied by Agents and is a subcategory of Agent Intrinsic Data.

(forall (x) (if (ActionExecutionTraceData x) (AgentIntrinsicData x)))

(forall (d r) (if (generates d r)
(and (ERAS-NEP:PlanAction d)
(ActionExecutionTraceData r))))

(forall (a m) (if (and (executes a m)
(ERAS-NEP:Agent a)
(ERAS-NEP:PlanAction m))
(exists (e)
(and (ActionExecutionTraceData e)
(generates m e))))))

Data in an Action Execution Trace tracks one or more Agent Actions and an Agent Action is tracked by zero or more Action Execution Traces.

(forall (d r) (if (tracks d r)
(and (ActionExecutionTraceData d)
(ERAS-NEP:AgentAction r))))

(forall (t) (if (ActionExecutionTraceData t)
(≥ 1 (#{ p | (and (tracks t p) (ERAS-NEP:AgentAction p)) }))))

(forall (p) (if (ERAS-NEP:AgentAction p)
(≥ 0 (#{ t | (and (tracks t p) (ActionExecutionTraceData t)) }))))

(forall (a p m e) (if (and (Agent a)
(PlanAction p)
(AgentAction m)
(ActionExecutionTraceDate e)
(is_implemented_by p m)
(executes a p))
(and (generates p e)
(tracks e m))))

As Autonomous Agents interact with other agents in a Social Interaction Process, an ordered sequence of events is emitted. The events in the interaction process are collected to provide a history of agent communications with other agents. This event history is classified as an Agent Interaction Trace Data and contains the sequence of events generated from agent interactions, communications, and agent plan executions.

The Agent Interaction Trace Data is a subcategory of Agent Intrinsic Data.

(forall (x) (if (AgentInteractionTraceData x) (AgentIntrinsicData x)))

(forall (x) (if (AgentInteractionTraceData x)
 (exists(d)
 (and (ERAS-TLO:SocialInteractionProcess d)
 (emits d x))))))

(forall (d r) (if (emits d r)
 (and (ERAS-TLO:SocialInteractionProcess d)
 (AgentInteractionTraceData r))))

A Social Interaction Process emits zero or more Agent Interaction Traces, and an Agent Interaction Trace is emitted by one or more Social Interaction Processes.

(forall (p) (if (ERAS-TLO:SocialInteractionProcess p)
 (>= 0 (#{ t | (and (emits p t) (AgentInteractionTraceData t)) }))))

(forall (t) (if (AgentInteractionTraceData t)
 (>= 1 (#{ p | (and (emits p t) (ERAS-TLO:SocialInteractionProcess p)) }))))

The Agent Static Data is a subcategory of Agent Intrinsic Data and classifies static continuant information about an Agent.

Autonomous Agents can have internal and external documentation and manuals with information about usage procedures, principles of operations, and the design, implementation, test and verification metrics collected during the system development life cycle. This information is available as Discourse Content for Agent Explanations and is represented as Agent Static Data.

(forall (x) (if (AgentStaticData x) (AgentIntrinsicData x)))

Agent Static Data has a range of sources that comprise it. A descriptive characterization of such source material includes the following enumerated examples:

```
(= (Description agent_documentation)
  { user_manual
    design_specifications
    principles_of_operations
    test_plans
    verification_metrics
  } )
```

Depending on the context of the Explanation request, some external information not directly affiliated with the Agent that is accountable for the Explanation may be accessed in the formulation of the Explanation's Discourse Content. The Agent Extrinsic Data category classifies this information.

Agent Extrinsic Data has a range of external information that comprise it. A descriptive characterization of such external information includes the following enumerated examples:

```
(= (Description external_information)
  { weather_forecast
    libraries
    news_media
    newspapers
    public_records
  } )
```

```
linked_data
agency_databases
MAS_collaboration
KB_inferencing
} )
```

The Content Provenance category classifies metadata information about the other five subcategories of Agent Data available for the formulation of Explanation Discourse Content. The Content Provenance data provides information about the Agents and Processes involved in the generation and composition of the respective sources of Agent Data formulated as Discourse Content. This information can be used to assess the quality, reliability and trustworthiness of the subject data provided in an Agent's Explanation.

```
(forall (d r) (if (authenticates d r)
  (and (ContentProvenance d)
    (DiscourseContent r))))
```

The Content Provenance concept provides this authentication metadata by asserting one or more Provenance Facts. Each Provenance Fact documents the author, the provider or authorizing agent, the generation or rendering process used, and the time of composition or editing for each referenced Agent Data source in the Discourse Content.

```
(forall (x) (if (ProvenanceFact x) (ERAS-TLO:InformationArtifact x)))
```

```
(forall (p) (if (ContentProvenance p)
  (exists (r)
    (and (ProvenanceFact r)
      (asserted_by r p))))))
```

```
(forall (d r) (if (asserted_by d r)
  (and (ProvenanceFact d)
    (ContentProvenance r))))
```

```
(forall (p) (if (ContentProvenance p)
  (>= 1 (#{ pf | (and (asserted_by pf p) (ProvenanceFact pf)) } ))))
```

```
(forall (d r) (if (is_metadata_about d r)
  (and (ContentProvenance d)
    (AgentData r))))
```

```
(forall (d r) (if (artifact d r)
  (and (ProvenanceFact d)
    (ERAS-TLO:InformationArtifact r))))
```

```
(forall (d r) (if (author d r)
  (and (ERAS-TLO:InformationArtifact d)
    (ERAS-TLO:Agent r))))
```

```
(forall (d r) (if (provider d r)
  (and (ERAS-TLO:InformationArtifact d)
    (ERAS-TLO:Agent r))))
```

(forall (d r) (if (renderer d r)
 (and (ERAS-TLO:InformationArtifact d)
 (ERAS-TLO:Process r))))

(forall (d r) (if (rendered_time d r)
 (and (ERAS-TLO:InformationArtifact d)
 (ERAS-TLO:Time r))))

The Provenance Fact concept asserts the author, the provider, the rendering process, and the rendered time associated with the source information artifact accessed for an Explanation's Discourse Content. Some of the information asserted by a Provenance Fact may have been derived from other Provenance Fact instances.

(forall (d r) (if (derived_from d r)
 (and (ProvenanceFact d)
 (ProvenanceFact r)
 (not (= d r))))))

Provenance metadata for the information artifact asserted by Provenance Facts include the author, provider, and rendering processes of the information artifact.

(forall (f s) (if (and (artifact f s)
 (ProvenanceFact f)
 (ERAS-TLO:InformationArtifact s))
 (exists (au p r t)
 (and (ERAS-TLO:Agent au)
 (ERAS-TLO:Agent p)
 (ERAS-TLO:Process r)
 (ERAS-TLO:Time t)
 (author s au)
 (provider s p)
 (renderer s r)
 (rendered_time s t))))))

(forall (x s cp dc ad au p r t)
 (if (and (ProvenanceFact x)
 (ERAS-TLO:InformationArtifact s)
 (ContentProvenance cp)
 (DiscourseContent dc)
 (AgentData ad)
 (ERAS-TLO:Agent au)
 (ERAS-TLO:Agent p)
 (ERAS-TLO:Process r)
 (ERAS-TLO:Time t)
 (artifact x s)
 (author s au)
 (provider s p)
 (renderer s r)
 (rendered_time s t))

```
(asserted_by x cp)  
(is_metadata_about cp ad)  
(accesses dc ad)  
(and (authenticates cp dc)  
(author ad au)  
(provider ad p)  
(renderer ad r)  
(rendered_time ad t))))
```

4.8 Ethical Violation Management

Figure 5 shows the UML diagram that contains the main concepts and relationships elicited during the investigation of the Ethical Violation Management (EVM) subdomain. It focuses on concepts and relationships associated with capabilities to detect, assess, and manage ethical violations in autonomous system behavior. In addition to ethical violation conceptualizations, this model also includes concepts and relationships governing accountability, responsibility, and legal notions of personhood for agents. Some concepts identified in the other subdomains are identified and shown in a specific color. An explanation about each concept is expressed in a mix of sentences in natural language and axioms.

During an agent interaction with the environment and other agents, some norms can be violated. A norm violation is an Action Event reflecting an agent's failure to conform to the norm's rules of behavior relevant to the agent's Situation.

```
(forall (x) (if (NormViolation x) (ERAS-TLO:ActionEvent x)))
```

```
(forall (d r) (if (is_violation_of d r)  
(and (NormViolation d)  
(ERAS-NEP:Norm r))))
```

A Norm Violation is a violation of one or more Norms.

```
(forall (v) (if (NormViolation v)  
(exist(n)  
(and (ERAS-NEP:Norm n)  
(is_violation_of v n))))))
```

```
(forall(v)(if (NormViolation v)  
(>= 1 (#{ n | (and (is_violation_of v n) (ERAS-NEP:Norm n)) } ))))
```

This violation might be detected by an Ethical Behavior Monitor, which is an agent or agent system component, either internal or external, that monitors AI Systems for normative ethical behavior conformance. The Norm Violation may also be documented and recorded in a Norm Violation Incident Information Artifact.

```
(forall (x) (if (EthicalBehaviorMonitor x) (ERAS-TLO:Object x)))
```

```
(forall (d) (if (EthicalBehaviorMonitor d)  
(exists(r)  
(and (ERAS-NEP:Agent r)  
(observes d r))))))
```

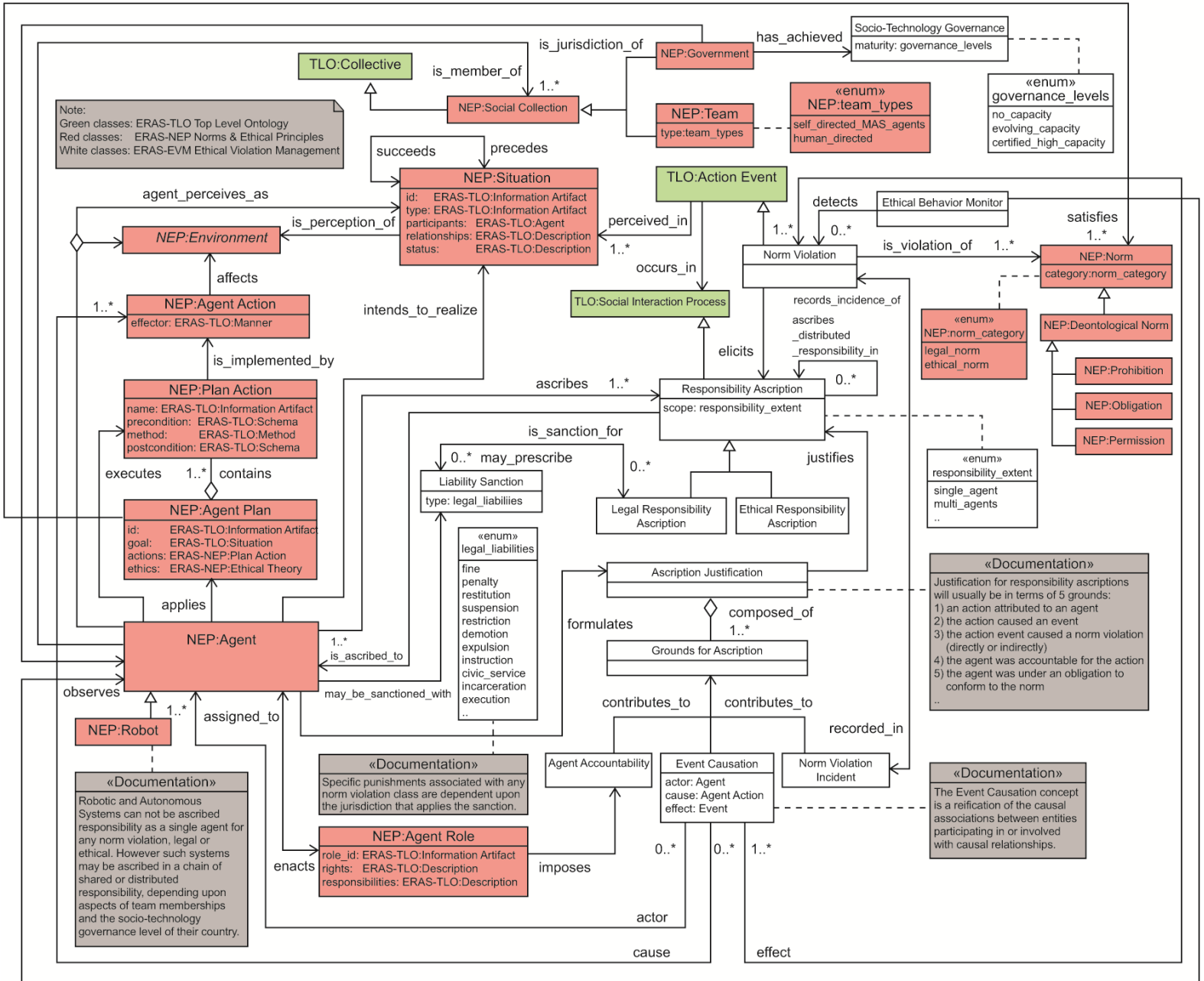


Figure 5 — Ethical Violation Management UML Diagram

(forall (d r) (if (observes d r)
 (and (EthicalBehaviorMonitor d)
 (ERAS-NEP:Agent r))))

(forall (d r) (if (detects d r)
 (and (EthicalBehaviorMonitor d)
 (NormViolation r))))

An Ethical Behavior Monitor may detect zero or more Norm Violations.

(forall (m) (if (EthicalBehaviorMonitor m)
 (>= 0 (# { v | (and (detects m v) (NormViolation v) }))))

(forall (x) (if (NormViolationIncident x) (ERAS-TLO:InformationArtifact x)))

(forall (d r) (if (records_incidence_of d r)
 (and (NormViolationIncident d)
 (NormViolation r))))

(forall (d r) (if (recorded_in d r)
 (and (NormViolation d)
 (NormViolationIncident r))))

(forall (nv) (if (NormViolation nv)
 (exists (evm n nvi)
 (and (EthicalBehaviorMonitor evm)
 (ERAS-NEP:Norm n)
 (detects evm nv)
 (is_violation_of nv n)
 (NormViolationIncident nvi)
 (recorded_in nv nvi)
 (records_incident_of nvi nv))))))

A Norm Violation elicits a Responsibility Ascription process as a Social Interaction Process to identify those responsible for the violation.

The Responsibility Ascription category classifies the process of assigning responsibility for a norm violation to an agent by an agent or agency acting in an authoritative role, either explicitly or implicitly. The responsibility ascription is justified by factual, legal, or ethical grounds that account for ascribing the subject agent as the actor responsible for the consequences of the action or actions causing the norm violation.

(forall (x) (if (ResponsibilityAscription x) (ERAS-TLO:SocialInteractionProcess x)))

(forall (d) (if (NormViolation d)
 (exists (r)
 (and (ResponsibilityAscription r)
 (elicits d r))))))

(forall (d r) (if (elicits d r)

(and (NormViolation d)
(ResponsibilityAscription r))))

A Responsibility Ascription is ascribed to at least one Agent by at least one Agent.

(forall (d r) (if (ascribes d r)
(and (ERAS-NEP:Agent d)
(ResponsibilityAscription r))))

(forall (r) (if (ResponsibilityAscription r)
(exists (a)
(and (ERAS-NEP:Agent a)
(ascribes a r))))))

(forall (r) (if (ResponsibilityAscription r)
(>= 1 (#{ a | (and (ascribes a r) (ERAS-NEP:Agent a) }))))

(forall (d r) (if (is_ascribed_to d r)
(and (ResponsibilityAscription d)
(ERAS-NEP:Agent r))))

(forall (r) (if (ResponsibilityAscription r)
(exists (a)
(and (ERAS-NEP:Agent a)
(is_ascribed_to r a))))))

(forall (r) (if (ResponsibilityAscription r)
(>= 1 (#{ a | (and (is_ascribed_to r a) (ERAS-NEP:Agent a) }))))

When ascribing responsibility for a Norm Violation, the ascribing Agent or Agency will need to consider the responsibility extent or scope of the participating agents associated with the norm violation. A descriptive characterization of a Responsibility Ascription scope includes the following examples:

(= (Description responsibility_extent)
{ single_agent
multi_agents
}))

When multiple agents are involved with a Norm Violation, ascription may entail a distributed responsibility relationship with separate justification grounds for each related Responsibility Ascription.

(forall (d r) (if (ascribes_distributed_responsibility_in d r)
(and (ResponsibilityAscription d)
(ResponsibilityAscription r)
(not (= d r))))))

(forall (r) (if (ResponsibilityAscription r)
(>= 0 (#{ d | (and (ascribes_distributed_responsibility_in r d)(ResponsibilityAscription d))}))))

The Responsibility Ascription category accounts for the two types of Norms, ethical and legal, with two subcategories: Ethical Responsibility Ascription and Legal Responsibility Ascription.

The Ethical Responsibility Ascription category classifies the process of assigning responsibility for an ethical norm violation.

(forall (x) (if (EthicalResponsibilityAscription x) (ResponsibilityAscription x)))

The Legal Responsibility Ascription category classifies the process of assigning responsibility for a legal norm violation.

(forall (x) (if (LegalResponsibilityAscription x) (ResponsibilityAscription x)))

A Legal Responsibility Ascription may prescribe zero or more Liability Sanctions. A Liability Sanction is a permitted punishment or penalty defined by an authoritative agency as a liability that may be imposed against an agent that is ascribed as responsible for a legal norm violation.

(forall (x) (if (LiabilitySanction x) (ERAS-TLO:Method x)))

(forall (d r) (if (is_sanction_for d r)
 (and (LiabilitySanction d)
 (LegalResponsibilityAscription r))))

(forall (s) (if (LiabilitySanction s)
 (>= 0 (# { r | (and (is_sanction_for s r)(LegalResponsibilityAscription r) }))))

(forall (d r) (if (may_prescribe d r)
 (and (LegalResponsibilityAscription d)
 (LiabilitySanction r))))

(forall (r) (if (LegalResponsibilityAscription r)
 (>= 0 (# { s | (and (may_prescribe r s)(LiabilitySanction s) }))))

(forall (d r) (if (may_be_sanctioned_with d r)
 (and (ERAS-NEP:Agent d)
 (LiabilitySanction r))))

(forall (a r) (if (and (ERAS-NEP:Agent a)
 (LegalResponsibilityAscription r)
 (is_ascribed_to r a)
 (exists (s)
 (and (LiabilitySanction s)
 (is_sanction_for s r)
 (may_be_sanctioned_with a s))))

The range of legal liabilities available as sanctions for specific Legal Norm Violations is dependent upon the jurisdiction that applies the sanction. A descriptive characterization of such sanctions includes the following examples:

(= (Description legal_liabilities)
 { fine
 penalty

restitution
suspension
restriction
demotion
expulsion
instruction
civic_service
incarceration
execution
)))

A Responsibility Ascription process that results in the ascription of responsibility to one or more Agents is justified by an Ascription Justification Information Artifact.

The Ascription Justification category classifies the collection of facts formulated and asserted by an authoritative agent or agency to ascribe responsibilities for ethical or legal Norm Violations. It is composed of constituent Grounds for Ascription Information Artifacts.

(forall (x) (if (AscriptionJustification x) (ERAS-TLO:InformationArtifact x)))

(forall (d r) (if (justifies d r)
 (and (AscriptionJustification d)
 (ResponsibilityAscription r))))

(forall (r) (if (ResponsibilityAscription r)
 (exists (a)
 (and (AscriptionJustification a)
 (justifies a r))))))

The Grounds for Ascription category classifies the collection of factual circumstances, causal events, and legal or ethical obligations that are evaluated to become the justification for ascribing responsibility for norm violations. Justification for Responsibility Ascription will usually be in terms of the following five grounds:

- d) An action attributed to an agent
- e) The action caused an event
- f) The action event caused a norm violation (directly or indirectly)
- g) The agent was accountable for the action
- h) The agent was under an obligation to conform to the norm

(forall (x) (if (GroundsForAscription x) (ERAS-TLO:InformationArtifact x)))

(forall (d r) (if (composed_of d r)
 (and (AscriptionJustification d)
 (GroundsForAscription r))))

(forall (j) (if (AscriptionJustification j)
 (exists (g)
 (and (GroundsForAscription g) (justifies g j))))))

(composed_of j g))))

(forall (j) (if (AscriptionJustification j)
(>= 1 (# { g | (and (composed_of j g) (GroundsForAscription g) })))))

The Agent Accountability category classifies the schemas that establish agent attributes such as age, physical and mental state, capabilities, intentions, knowledge, role responsibilities, and authority that contribute to the assessment of the agent's or agency's responsibility for a Norm Violation.

(forall (x) (if (AgentAccountability x) (ERAS-TLO:Schema x)))

Agent Roles enacted by Agents imposes Agent Accountability.

(forall (x) (if (AgentAccountability x)
(exists (a)
(and (ERAS-NEP:AgentRole a)
(imposes a x))))))

(forall (d r) (if (imposes d r)
(and (ERAS-NEP:AgentRole d)
(AgentAccountability r))))

As an actor or participant in an Interaction Process, an Agent may apply an Agent Action that causes one or more Action Events that manifest as Norm Violations. This EVM subdomain reifies these associations between entities participating in or involved with such causal relationships with the category of Event Causation.

The Event Causation category identifies an Agent actor, the Agent Action, and the Norm Violation Action Event effect that contribute to the assessment of the agent's or agency's responsibility for a Norm Violation.

(forall (x) (if (EventCausation x) (ERAS-TLO:InteractionProcess x)))

(forall (d r) (if (actor d r)
(and (EventCausation d)
(ERAS-NEP:Agent r))))

(forall (d r) (if (cause d r)
(and (EventCausation d)
(ERAS-NEP:AgentAction r))))

(forall (d r) (if (effect d r)
(and (EventCausation d)
(NormViolation r))))

An Event Causation process has at least one Agent actor. An Agent may be an actor in zero or more Event Causations.

(forall (x) (if (EventCausation x)
(exists (a)
(and (ERAS-NEP:Agent a)

(actor a x))))))

(forall (e) (if (EventCausation e)
(≥ 1 (# { a | (and (actor a e) (ERAS-NEP:Agent a) })))))

(forall (a) (if (Agent a)
(≥ 0 (# { e | (and (actor a e) (EventCausation e) })))))

An Event Causation process has at least one Agent Action in the causal chain.

(forall (e) (if (EventCausation e)
(exists (a)
(and (ERAS-NEP:AgentAction a)
(cause a e))))))

(forall (e) (if (EventCausation e)
(≥ 1 (# { a | (and (cause a e) (ERAS-NEP:AgentAction a) })))))

An Agent Action may have caused zero or many Event Causations.

(forall (a) (if (ERAS-NEP:AgentAction a)
(≥ 0 (# { e | (and (cause a e) (EventCausation e) })))))

An Event Causation process has at least one Norm Violation associated with the effect of the Agent Action.

(forall (e) (if (EventCausation e)
(exists (a)
(and (NormViolation a)
(effect e a))))))

(forall (e) (if (EventCausation e)
(≥ 1 (# { a | (and (effect e a) (NormViolation a) })))))

A Norm Violation event may be implicated in one or more Event Causations.

(forall (a) (if (NormViolation a)
(≥ 1 (# { e | (and (effect e a) (EventCausation e) })))))

An Event Causation entity identifies the Agent actor and Agent Action associated with a Norm Violation Action Event.

(forall (ec) (if (EventCausation ec)
(exists (aa c pa nv m)
(and (ERAS-NEP:Agent a)
(EthicalBehaviorMonitor m)
(ERAS-NEP:AgentAction aa)
(NormViolation nv)
(ERAS-NEP:PlanAction pa)
(is_implemented_by pa aa)
(executes a pa))))))

(observes m a)
(detects m nv)
(actor ec a)
(cause ec aa)
(effect ec nv))))))

The categories of Agent Accountability, Event Causation, and Norm Violation Incident may all contribute to the Grounds for Ascription that comprises Ascription Justification for a Responsibility Ascription.

(forall (g) (if (GroundsForAscription g)
 (exists (a e n)
 (and (AgentAccountability a)
 (EventCausation e)
 (NormViolationIncident n)
 (contributes_to a g)
 (contributes_to e g)
 (contributes_to n g))))))

(forall (d r) (if (contributes_to d r)
 (and (or (AgentAccountability d)
 (EventCausation d)
 (NormViolationIncident d))
 (GroundsForAscription r))))

As facts and evidence gathered in the Information Artifacts of Agent Accountability and Norm Violation Incidents are combined with the Event Causation relationships that may contribute to a Grounds for Ascription, an authorized Agent or Agency formulates the Ascription Justification to justify a Responsibility Ascription.

(forall (ra) (if (ResponsibilityAscription ra)
 (exists (a b aj)
 (and (ERAS-NEP:Agent b)
 (ERAS-NEP:Agent a)
 (AscriptionJustification aj)
 (not (= a b))
 (is_ascribed_to ra b)
 (ascribes a ra)
 (justifies aj ra)
 (formulates a aj))))))

Alternative World View axiom patterns regarding aspects of Distributed Responsibility Ascription for Autonomous Agents follow next.

The Robot category defined in the NEP subdomain is a synonym for Autonomous System and denotes both physically embodied and non-embodied AI systems.

During the EVM subdomain analysis regarding the extent to which ethically aware autonomous systems could be ascribed any degree of responsibility for a Norm Violation, predispositions aligning with three separate world views emerged.

A Legal World View (LWV) predisposition maintains that current and foreseeable future legal systems do not and should not permit ascribing responsibility to autonomous systems for any norm violation, legal or ethical. In the current LWV, human and other agencies granted personhood can be ascribed partial or distributed responsibility for participating in activities resulting in a Legal Norm violation, but autonomous systems cannot.

A Technology World View (TWV) predisposition maintains that emerging advances in AI technology will soon motivate granting autonomous systems with formal and legal agenthood with consequential accountability and responsibility requirements. Upon achieving the requisite advances such autonomous systems should be ascribed both direct and distributed responsibility for their actions.

A Common World View (CWV) proposes a middle ground between the LWV and the TWV. It is based on the concept of a maturity level of socio-technology governance capabilities to be achieved and certified for governments adopting the ERAS ontology commitments. The extent and type of responsibility ascriptions that can be ascribed to ethically aware autonomous systems would be based on the level of socio-technology governance achieved.

To formalize the CWV commitments, the Ethical Violation Management (EVM) subdomain proposes three axiom pattern sets that correspond to three qualitative maturity levels of Socio-Technology Governance capabilities to which Governments may aspire and achieve as their respective laws and social customs evolve to accommodate technological advances in autonomous robotic systems.

A government's level of socio-technology governance capabilities will depend upon implementation of social and legal requirements deemed necessary for granting autonomous AI systems some notion of legal personhood. As described in van Generen [B54] and Pagallo [B42], such requirements will include the following:

- Necessity in the “human” society for legal certification.
- Acceptance by other legal persons by creating trust and reliance for other legal and natural persons to integrate in economic, social, and legal interactions.
- Sufficient social intelligence on the part of AI systems with capabilities to understand the socio-emotional and moral value of statements by other parties.
- Adaptive capabilities on the part of AI systems to respond to changing circumstances.
- A public register that specifies which AI systems will have specific legal competences for specified roles and tasks.
- The creation of special zones for AI and robotics empirical testing and development.
- The conduction of experiments to determine system safety and relevant legal liabilities.
- Exploration of new forms of accountability and liability with a focus on complex distributed responsibility and strategies such as obligatory insurance policies.
- Establishing insurance systems for producers and providers supplemented with funds to insure compensation of damages.

The category of Socio-technology Governance classifies the endeavors of government agencies to provide oversight and management of the intersecting social and technological processes that create, modify, and sustain the design and introduction of artifacts and methods involved in complex systems that entail aspects of both technological and sociological systems.

(forall (x) (if (Socio-TechnologyGovernance x) (ERAS-TLO:InteractionProcess x)))

Over time, governments may achieve and be certified to have achieved required levels of Socio-Technology Governance maturity.

(forall (d r) (if (has_achieved d r)
(and (ERAS-NEP:Government d)
(Socio-TechnologyGovernance r))))

A descriptive characterization of possible qualitative evaluations gaging the level of capacity that would distinguish the capabilities of a country's socio-technology governance includes the following examples:

(= (Description governance_levels)
{ no_capacity
evolving_capacity
certified_high_capacity
})

(forall (g mg) (if (and (ERAS-NEP:Government g)
(Socio-TechnologyGovernance mg)
(has_achieved g mg))
(or (= (maturity mg) no_capacity)
(= (maturity mg) evolving_capacity)
(= (maturity mg) certified_high_capacity))))

The three axiom patterns that correspond to the three maturity levels for a Government's Socio-Technology Governance capabilities are as follows:¹²

- a) *Axiom Pattern A for Governments that have no capacity:* An Autonomous System cannot be ascribed responsibility for any Norm Violation, ethical or legal, either as a single actor of an Event Causation, or as a team member of a human directed team of actors associated with an Event Causation. This makes the Common World View (CWV) equivalent to the Legal World View (LWV) currently in effect throughout the world.
- b) *Axiom Pattern B for Governments achieving an evolving capacity:* An Autonomous System can only be ascribed distributed responsibility for an Ethical Norm Violation.
- c) *Axiom Pattern C for Governments achieving a certified high capacity:* An Autonomous Robotic System cannot be ascribed responsibility as a single agent for any norm violation, legal or ethical. However, an Autonomous Robotic system may be encumbered with a distributed responsibility ascription as a member of a multi-agent team directed by a human agent if the Government in which the system is being ascribed as responsible has achieved a certified high capacity level for their Socio-Technology Governance policies.

Subclauses 4.8.1 and 4.8.2 present axioms definitions for pattern A, where governments have no capacity, and for pattern B, where governments have an evolving capacity. However, since currently there are no cases in which a government has a certified high level of Socio-Technology Governance, a majority of the P7007 contributors voted to place the axiom definitions for pattern C in Annex D. The objective of defining these axioms as informative instead of normative is to invite and motivate discussion across the stakeholder communities.

¹² The Capability Maturity Model (CMM®) developed by the Software Engineering Institute at Carnegie Mellon University is a similar organizational maturity level evaluation framework and provides analogous evaluation examples for the socio-technology governance maturity levels involved with the preconditions listed in van Genderen [B54]. Caputo [B17] provides more information on CMM.

4.8.1 Axiom Pattern A for Governments with no capacity

When a Government has a “no capacity” maturity level for its Socio-technology Governance, then an Autonomous System cannot be ascribed responsibility for any Norm Violation, ethical or legal, either as a single actor of an Event Causation, or as a team member of a human directed team of actors associated with an Event Causation.

```
(forall (ra g mg) (if (and (ERAS-NEP:Government g)
                          (Socio-TechnologyGovernance mg)
                          (= ( maturity mg) no_capacity)
                          (has_achieved g mg)
                          (ERAS-NEP:Agent a)
                          (is_jurisdiction_of g a)
                          (ResponsibilityAscription ra)
                          (is_ascribed_to ra a))
                      (not (ERAS-NEP:Robot a))))
```

This is equivalent to the strict legal world view which contemplates no differentiation regarding Socio-Technology Governance.

```
(forall (ra a) (if (and (ResponsibilityAscription ra)
                       (is_ascribed_to ra a)
                       (ERAS-NEP:Agent a))
                  (not (ERAS-NEP:Robot a))))
```

4.8.2 Axiom Pattern B for Governments achieving an evolving capacity

When a Government has an “evolving capacity” maturity level for its Socio-technology Governance then an Autonomous System can only be ascribed distributed responsibility for an Ethical Norm Violation when the system was a member of a multi-agent team directed by a human and the Norm Violation was caused by an action of the autonomous system.

Distributed Ascriptions involve multiple agents.

```
(forall ( ra da ) (if (and (ResponsibilityAscription ra)
                          (ResponsibilityAscription da)
                          (not (= ra da))
                          (ascribes_distributed_responsibility_in ra da))
                      (= (scope ra) multi_agents )))
```

```
(forall ( h r hrt hra g mg ga aj n nv ec eda)
  (if (and (ERAS-DPP:Person h)
          (ERAS-NEP:Robot r)
          (ERAS-NEP:Team hrt)
          (= (type hrt) human_directed)
          (is_member_of h hrt)
          (is_member_of r hrt)
          (ResponsibilityAscription hra)
          (= (scope hra) multi_agents)
          (is_ascribed_to hra h)
          (ERAS-NEP:Government g)
          (is_jurisdiction_of g h)
          (is_jurisdiction_of g r))
```

(Socio-TechnologyGovernance mg)
(= (maturity mg) evolving_capacity)
(has_achieved g mg)
(NormViolation nv)
(ERAS-NEP:Norm n)
(= (category n) ethical_norm)
(is_violation_of nv n)
(EventCausation ec)
(= (actor ec) r)
(= (effect ec) nv)
(GroundsForAscription ga)
(contributes_to ec ga)
(AscriptionJustification aj)
(composed_of aj ga)
(EthicalResponsibilityAscription eda)
(justifies aj eda))
(and (is_ascribed_to eda r)
(ascribes_distributed_responsibility_in eda hra))))

Annex A

(informative)

Informative definitions

This annex presents the informative definitions for the concepts and items of the enumerated sets in alphabetic order grouped according to the ontology where they appear. The corresponding formal definitions are already presented across the main body of this standard through axioms expressed in CLIF.

A.1 Top-level definitions

abstract: An entity subcategory that classifies non-physical conceptualizations that have no locations in space or time.

action event: An Event subcategory that classifies Event occurrences generated by Agents within a process.

agent: An Object subcategory that classifies physical entities that can act autonomously and produce changes in their situated environment.

agent communication: An Action Event subcategory that classifies an Action Event occurrence generated by Agents to transmit information within a process.

attribute: A Property subcategory that classifies entities which are properties of some Continuant object.

collective: An Abstract subcategory that classifies entities which are grouped together according to some constitution relation. Entities that are members of a collection have uniform structure.

continuant: A Physical subcategory that classifies physical entities with stable attributes or characteristics that enable them to be recognized as the same individual or instance over time.

description: An Abstract subcategory that classifies entities that specify aspects or characteristics of other physical or abstract entities.

end time: An Event category property designating the ending time point of the Event.

entity: The universal, top-level category in the ERAS:TLO taxonomy of concepts. All ERAS ontology concepts are subcategories of Entity and all instances of those categories are instances of Entity.

environmental event: An Event subcategory that classifies Event occurrences generated by non-agent processes.

event: An Occurrent subcategory that classifies entities with initiation and termination time points occurring within a Physical Process.

information artifact: An Object subcategory that classifies entities that render abstract descriptive ideas, expressions, and facts as tangible artifacts using printed text, electronic media, or some form of physical substrate.

interaction process: A Process subcategory that classifies Process occurrences that includes at least one Action Event that was generated by one or more Agents.

manner: A Property subcategory that classifies entities which are properties of some Occurrent process.

method: A Description subcategory that classifies entities which are abstract descriptions of Occurrent Process actions to produce some result.

object: A Continuant subcategory that classifies entities which retain their identity over time and which can be perceived when observed as complete instances. Object entities can have different properties at different times and therefore can undergo change.

occurrent: A Physical subcategory that classifies physically occurring entities that do not have a stable identity during an interval of time. Occurrent entities may have phases that extend in time but that are not wholly perceived at any point in time.

physical: An entity subcategory that classifies entities which have a location in space-time, that is those entities that are located within a specific space at a specific time.

plan: A Method subcategory that classifies entities which specify a sequence of processes intended to satisfy a specified purpose or goal for an Agent by affecting changes in the Agent's physical situation.

process: An Occurrent subcategory that classifies entities that last in time but which can only be partially perceived when observed at a specified time. A Process entity is not an object but may have participants within it that are objects.

property: An Abstract subcategory that classifies entities which characterize features of entities perceived by Agents and which are distinguished by the category of bearing entities as qualities of Continuant entities or qualities of Occurrent entities.

role: An Occurrent subcategory that classifies entities which specify permissions, obligations, and relational aspects for Agents that enact the role.

schema: A Description subcategory that classifies entities which are abstract descriptions of configuration or structural aspects of Continuant entities.

situation: An Occurrent subcategory that classifies aggregated instances comprised of participating entities and relationships among them which represent the limited parts of reality that can be perceived and reasoned about by agents.

social interaction process: An Interaction Process subcategory that classifies Interaction Process occurrences that includes multiple Agents engaged in Agent Communication sub process stages.

spatio temporal place: An Abstract subcategory that classifies the spatial and temporal qualities representing the places in which Physical entities are located at specific times.

start time: An Event category property designating the initiating time point of the Event.

time: An Abstract subcategory that classifies a linear sequence of time points. Points within such a sequence are time values at which Physical entities may be located within a spatio-temporal place at that time.

A.2 Norms and Ethical Principles

action rationale: A method (TLO:Method) subcategory that logically justifies a Plan Action as an appropriate Agent autonomous action. Logical justifications for a Plan Action are based on autonomous action principles such as informed consent, autonomy and unethical prevention interference.

activated: A Norm State denoting Norm entities that are active and influencing the selection of Plan Actions in Agent Plans.

agent action: A process (TLO:Process) that is an operation or effector that implements the method specified in the plan action. The agent action is applied and executed by the agent to affect state changes in an agent's situated environment.

agent plan: An Information Artifact (TLO:InformationArtifact) subcategory that consists of specifications, partial or complete, for a sequence of agent actions to achieve target goals, objectives, and services to realize agent intentions. Agent Plans as subclasses of ERAS-TLO:InformationArtifact render ERAS-TLO:Abstract Plans into some physical substrate.

agent role: A role (TLO:Role) subcategory that characterizes a defined set of connected behaviors, capabilities, requirements, rights, obligations and permissions expected of any agent assigned or ascribed the respective agent role.

agent: An Agent subcategory (TLO:Agent) that classifies a Continuant Object entity that can act autonomously and produce changes in its situated environment.

answer: An Agent Communication (NEP:AgentCommunication) subcategory that classifies an Action Event responding with information that answers prior queries. The response may be expressed informally in natural language, formally using some formal query language, or using some visual medium.

autonomous action principles: An Action Rationale property range type that characterizes ethical principles that can explain and justify selected actions as rational and coherent.

autonomy: An Ethical Principle category type that classifies Norm accommodation and obligation for acts that respect the independency, sense of self, and capacity of agents to determine their own destiny.

beneficence: An Ethical Principle category type that classifies Norm accommodation and obligation based on acts that benefit others with behavior exhibiting conditions of charity, mercy, kindness, and moral imperatives.

buddhist ethics: An Ethical Theory modality that classifies behavior that promotes an Agent's correct action to be based on Buddhist practices such as commitment to harmony, nonviolence, respect, security, virtuous obligations, and causing no harm.

civility: An Ethical Principle category type that classifies Norm accommodation and obligation for acts and social interactions that exhibit formal politeness, courtesy, and etiquette.

community: A social collection (NEP:SocialCollection) subcategory that is an aggregation of agents grouped together by common properties such as geographic location, ethnic affiliations, or shared values.

company: A social collection (NEP:SocialCollection) subcategory that is an aggregation of agents as employees of a company.

composite: An Ethical Theory modality that classifies behavior that promotes an Agent's correct action to be based on an aggregation of many Ethical Theories where the Agent's situated environment determines which Ethical Theory best influences action choices.

consequentialist norm: A norm (NEP:Norm) subcategory derived from the Consequentialist ethical theory that elucidates correct action choices based on the consequences that the action produces. Generally, actions that are expected to result in a greater intrinsic good are to be preferred.

consequentialist: An Ethical Theory modality that classifies behavior that promotes an Agent's correct action to be based on the consequences that the action produces.

deontological norm: A norm (NEP:Norm) subcategory derived from the deontological ethical theory that stipulates correct action choices based on the action's conformity to universal rules for judging rightness or wrongness of an act. From this perspective, correct behavior is independent of the resulting consequences.

deontological: An Ethical Theory modality that classifies behavior that promotes an Agent's correct action to be based on the action's conformity to universal rules for judging rightness or wrongness.

department: A social collection (NEP:SocialCollection) subcategory that is an aggregation of agents belonging to a subgroup that is part of a larger group, company, or organization.

derogation: A process (TLO:Process) subcategory that an agent activates for the purpose of temporarily suspending or derogating a norm

dilemma mitigation principle: A method subcategory (TLO:Method) that specify descriptions of action properties or conditions that should hold in order that an action may be deemed morally permissible when applied to resolve an ethical dilemma. Examples include the principle of double effect and the principle of triple effect.

dilemma mitigation principles: A Dilemma Mitigation Principle property range category that classifies the type of principles available for resolving Ethical Dilemma conflicts by ranking a preferred Norm.

divine command: An Ethical Theory modality that classifies behavior that promotes an Agent's correct action to be based on a rigorous logical specification of rules collectively deemed to be morally and ethically applicable for autonomous systems and where behaviors outside the specified rules are not permitted.

environment: A continuant (TLO:Continuant) subcategory that classifies an external collection of entities, entity properties, entity relationships, and occurrent processes that pose potential internal Agent conceptualizations derived from Agent perceptions of the external entities present in the Environment.

ethical dilemma: A situation subcategory (TLO:Situation) arising between conflicting normative rules of agent behavior in which none of the choices are deemed unambiguously acceptable or preferable.

ethical norm: A Norm Category type that classifies a Norm entity as an ethical norm that pertains to a general rule of behavior expected by society and not necessarily enforced as a principle of law.

ethical Principle: A Method (TLO:Method) subcategory that identifies principles of agent behavior in terms of moral proposition and value judgements which characterize and justify particular ethical prescriptions and evaluations of agent actions.

ethical Principles: An Ethical Principle property range category that classifies the type of Ethical Principle that is accommodated by Norms and the Norm's Ethical Theory modality.

ethical theories: An Ethical Theory property range category that classifies the type of Ethical Theory modality that constrains plans for the Agent's Situation Plan Repertoire.

ethical theory: A method (TLO:Method) subcategory that is a systematization of concepts specifying or recommending aspects of morally correct behavior based on philosophical values and the characterization of right and wrong conduct. For norm aware agents, normative ethical theory is concerned with the practical means of determining a moral course of action.

expired: A Norm State denoting Norm entities that are no longer applicable for Agent Plans and associated Plan Actions.

explanation: An Agent Communication (NEP:AgentCommunication) subcategory that classifies Action Events that respond to a request to explain and justify system behavior. The response may be tailored to the type and role of the agent making the request. (See an expanded informative definition of the Explanation category in the context of the TA subdomain context).

fairness: An Ethical Principle category type that classifies Norm accommodation and obligation for acts that demonstrate impartial and just treatment without favoritism or discrimination.

fidelity: An Ethical Principle category type that classifies Norm accommodation and obligation for acts that demonstrate faithfulness to norms, goals, missions, agents, and teams.

fulfilled: A Norm State denoting Norm entities that were satisfied by Agent Plans as an Agent applied its selected Agent Plans and executed the constituent Plan Actions.

generalization principle: An autonomous action principle that justifies an Agent Action as rational when the reasons for the action are consistent with the assumption that all agents with the same reasons take the same action.

government: A social collection (NEP:SocialCollection) subcategory that is an aggregation of agents participating in a governmental system that governs an organized community or state for the purpose of establishing direction, rights, obligations and control over members of the community or state.

human directed: A Team category type that classifies Social Collection Teams comprised of humans and autonomous agents systems that are directed by at least one human.

interference principle: An autonomous action principle that justifies an Agent Action as rational and coherent when the Agent selects an action that interferes with unethical action plans of other agents. An Agent Action that interferes with unethical action plans of another Agent does not compromise the autonomy of the Agent contemplating or engaging in unethical behavior.

justice: An Ethical Principle category type that classifies Norm accommodation and obligation for acts that exhibit the quality of being just, equitable and morally right or restorative.

legal norm: A Norm Category type that classifies a Norm entity as a legal norm that pertains to a binding rule or principle of law.

nonmaleficence: An Ethical Principle category type that classifies Norm accommodation and obligation for acts that avoid causing harm.

norm category: A Norm property range category type that defines a Norm entity as a legal norm or as an ethical norm.

norm states: A Norm property range category that classifies the life cycle state of Norm entities as they become associated with and satisfied by Agent Plans.

norm: A method (TLO:Method) subcategory that describes a set of rules and methods governing expected behavior for norm-aware agents.

not applicable: A Norm State denoting Norm entities that are not applicable for an Agent Plan selected by an Agent in its situated Environment.

obligation: A deontological norm subcategory (NEP:DeontologicalNorm) that specifies what an agent should do. An attribute that applies to propositions that an agent is required by some authority to make true.

organization: A social collection (NEP:SocialCollection) subcategory that is an aggregation of agents belonging to a group of participants with a shared purpose.

permission: A deontological norm subcategory (NEP:DeontologicalNorm) that specifies what an agent may do. An attribute that applies to propositions that an agent is permitted, by some authority to make true.

plan action: A method (TLO:Method) subcategory that is constituent of an agent plan and specifies the preconditions and postconditions for the application of an agent action to achieve the objectives and goals of the plan.

principle of autonomy: An autonomous action principle that justifies an Agent Action as rational and coherent if the Agent believes the action will not interfere with the action plans of another agent.

principle of double effect: A Dilemma Mitigation Principle category type that classifies a reasoning strategy for resolving ethical dilemmas by evaluating an act which entails foreseen harmful effects as permissible if a) the act will lead to a greater good, and b) the bad effect is an unintended consequence, and c) the bad effect is not the means of achieving the greater good.

principle of informed consent: An autonomous action principle that justifies an Agent Action as rational and coherent when the Agent believes that although the action may interfere with the action plans of another agent, that agent has given informed consent to the possibility of interference, and that the giving of that consent is itself a coherent action plan.

principle of triple effect: A Dilemma Mitigation Principle category type that classifies a reasoning strategy that refines the Principle of Double Effect strategy for resolving ethical dilemmas by deeming an action as permissible if it achieves a greater good by directly causing a bad effect but when the bad effect was not the intended goal of the agent.

prohibition: A deontological norm subcategory (NEP:DeontologicalNorm) that specifies what an agent is forbidden to do. An attribute that applies to propositions that an agent is forbidden, by some authority to make true.

query: An Agent Communication (NEP:AgentCommunication) subcategory that classifies an Action Event from an Agent requesting information about some topic. The inquiry may be expressed informally in natural language, formally using some formal query language, or using some visual medium.

respect: An Ethical Principle category type that classifies Norm accommodation and obligation for acts that demonstrate admiration, esteem, and consideration towards individuals based on qualities, achievements or status of the individual.

robot: A subcategory of Agent (NEP:Agent) that is also equivalent to the CORA:Robot concept formalization in the IEEE Std 1872-2015 Standard. The ERAS and equivalent CORA conceptualization denotes an agentive system provisioned with suitable components that enable the system to act in its physical environment to accomplish tasks.

self directed mas agents: A Team category type that classifies Social Collection Teams comprised of self directed autonomous agents systems or multi-agent systems.

situation plan repertoire: An information artifact (TLO:InformationArtifact) subcategory containing Agent Plans relevant for provisioning Agents with Agent Actions that enable achievement of agent intentions and behaviors appropriate for norm-aware agents.

situation: A physical continuant entity (TLO:Situation) characterizing internal Agent perceptions of the Agent's environment in which the Agent is situated. Norm-aware Agents perceive, recognize, and become aware of Situations presented in their environments.

social collection: A collective (TLO:Collective) subcategory that corresponds to an aggregation of agents grouped together by some common property or social purpose.

suspended: A Norm State denoting Norm entities that are temporarily inactivated by an Agent Action that activates a Derogation process.

task assignment: An Agent Communication (NEP:AgentCommunication) subcategory that classifies communication Action Events that assigns and specifies a mission, chore, duty, problem, or goal to undertake and accomplish or solve. The task specification may include initial conditions, a goal, assertions, and characterizations of available operations and resources, which are then represented in a task goal situation.

team types: A Team property range category that classifies the type of Team Social Collection entities.

team: A social collection (NEP:SocialCollection) subcategory that is an aggregation of agents formed for some usually short term objective.

veracity: An Ethical Principle category type that classifies Norm accommodation and obligation for acts that demonstrate truthfulness, accuracy, and correctness.

violated: A Norm State denoting Norm entities that were not satisfied by Agent Plans as an Agent applied its selected Agent Plans and executed the constituent Plan Actions.

virtuous norm: A norm (NEP:Norm) subcategory derived from the virtuous ethical theory that elucidates correct action choices based on alignment with certain dispositional character traits or virtues that are appropriate and praiseworthy. From this perspective, correct agent behavior is achieved by adhering to character traits deemed praiseworthy and not blameworthy.

virtuous: An Ethical Theory modality that classifies behavior that promotes an Agent's correct action to be aligned with dispositional character traits or virtues that are appropriate and praiseworthy.

A.3 Data Protection and Privacy

access policy: A method (TLO:Method) subcategory that is a data privacy and protection policy which specifies the requirements and prerequisites necessary to control and protect the collection, access, and use of personal data about the data subject.

access prerequisites: The prerequisites property range category of the Access Policy category. The prerequisite property of the Access Policy category identifies the subcategory of access prerequisites for instances of Access Policy.

accessible personal data: A personal data (DPP:PersonalData) subcategory which classifies that portion of data for which the data subject may grant consent for such access.

aggregated personal data: A personal data (DPP:PersonalData) subcategory that classifies data that has been collected, compiled, or data mined across multiple sources including public and private databases, social media, web sites and personal artifacts that can be used to infer and reveal new or previously unpublished and unavailable personal information about a data subject.

authorized accessor: A legislated governance role (DPP:LegislatedGovernanceRole) abstract subcategory representing common properties and relationships assigned to persons, natural or legal, public authorities or

agencies other than data subjects, and controllers that have been authorized by a controller to process personal data. Subclasses of this abstract concept represent the specific Accessor roles that may be assigned to agents.

auto GPS systems: The Environment Data subcategory classifying personal data associated with automotive Global Positioning Systems used by an individual. This information also includes the history of an individual's location data collected by such systems.

biometric data: The Health Data subcategory classifying physical or behavioral human characteristics that can be used to digitally identify a person. Examples include fingerprints, facial patterns, voice, eye iris patterns, DNA, and signatures.

care giver: An Agent Role category enacted by an Agent with the responsibility to provide care for another person that needs assistance and support. The Agent or Person receiving the care may be a child or minor, or they may be someone unable to manage their own affairs. A dependency relationship is established between the Agent enacting the Care Giver role and the Agent receiving the care.

care giver roles: A role property range type of the Care Giver category that classifies and distinguishes between various types of care giver roles.

care providers: The Health Data subcategory classifying personal data relating to the identification of professional providers of health care for individuals.

career: The Social Data subcategory classifying personal data associated with the identification of an individual's professional career choices and progression.

controller legal obligation: The Access Policy prerequisite subcategory permitting a Data Access Process collecting an individual's Accessible Personal Data when there is a controller legal obligation on the part of the controller that administers the Data Access Process.

controller sanctioned permission: The Access Policy prerequisite subcategory permitting a Data Access Process collecting an individual's Accessible Personal Data when there is a controller sanctioned permission on the part of the controller that administers the Data Access Process.

controller: A legislated governance role (DPP:LegislatedGovernanceRole) subcategory enacted by a natural or legal person, public agency or other body with the authority to determine, either alone or jointly, the purposes and means of processing personal data.

country: An Continuant (TLO: Continuant) subcategory that denotes a geographical territory in which physical objects may be located and which may be governed by a Government.

credit card data: The Economic Data subcategory classifying personal data associated with information generated by personal purchasing transactions using a credit card.

credit ratings: The Economic Data subcategory classifying personal data reflecting a quantified assessment of the creditworthiness of a person with respect to borrowing, debt, and financial obligation.

cross border transfer: The Data Access Process subcategory classifying Data Access Processes employed to apply rules for permitting or preventing transfer of a Person's Accessible Personal Data across international borders.

cultural: The Social Data subcategory classifying personal data associated with an individual's affinity to ideas, customs, and the social behaviors of a society.

data access form: The type property range category of the Data Access Process category. The type property of the Data Access Process category identifies the subcategory type for instances of Data Access Process.

data access process: A process (TLO:Process) subcategory that classifies processes comprised of a sequence of operations that have been authorized and validated by Agents enacting the relevant Legal Governance Roles in effect for the Person's circumstances.

data breach: A personal data transaction (DPP:PersonalDataTransaction) subcategory that classifies an incident in which sensitive, protected or confidential Personal Data is copied, transmitted, viewed, stolen, or used by an agent, or a personal data transaction on behalf of an agent, that is unauthorized to do so and for which the data subject has not granted consent to access.

data mining: The Data Access Process subcategory classifying Data Access Processes employed to discover and extract data about an individual by analyzing large and usually distributed sets of data.

data processor: An authorized accessor (DPP:AuthorizedAccessor) subcategory role enacted by a natural or legal person, public authority, agency, or other body which processes personal data as authorized by the controller.

data protection authority: A legislated governance role (DPP:LegislatedGovernanceRole) subcategory that defines the principal supervisory authority responsible for consistent application and enforcement of personal data and privacy protection policies and directives. A DPA becomes the main point of contact for participating stakeholder communities.

data protection by default: The Data Protection Principle subcategory that denotes technical and organizational methods that restrict personal data processing to access only that data required for the purpose of the Data Access Process and associated Personal Data Transactions. No additional personal data can be processed or made publicly available unless the individual consents to the transactions and grants access.

data protection by design: The Data Protection Principle subcategory that denotes technical and organizational practices to ensure the protection, privacy, and safeguarding of individual rights by integrating data protection features as essential core functions of systems and processes throughout all system lifecycle phases.

data protection principle: A method (TLO:Method) subcategory that articulates general guidelines intended to enable the protection and use of personal data across evolving technology and multiple stakeholder communities.

destruction: The Data Access Process subcategory classifying Data Access Processes employed to delete data from a Person's Accessible Personal Data.

dissemination: The Data Access Process subcategory classifying Data Access Processes employed to copy and distribute a Person's Accessible Personal Data.

econ data types: The type property range category of the Economic Data subcategory. The type property of Economic Data identifies the subcategory type for instances of Economic Data.

economic data: A personal data (DPP:PersonalData) subcategory that classifies information associated with the economic state and characteristics of the data subject.

economic: The Personal Data type value denoting the economic data subcategory classification of Personal Data.

education: The Social Data subcategory classifying personal data that identifies an individual's record of educational attainment including degrees, honors, and subject matter majors.

employment: The Economic Data subcategory classifying personal data associated with an individual's history of employment including information about employers, positions, skills, salaries, and staff evaluations.

entertainment systems: The Environment Data subcategory classifying personal data associated with an individual's installed home audio and video systems such as TVs, DVRs, AV Receivers, Surround Sound Speakers, and Electronic Game Consoles.

environment data types: The type property range category of the Environment Data subcategory. The type property of Environment Data identifies the subcategory type for instances of Environment Data.

environment data: A personal data (DPP:PersonalData) subcategory that classifies data associated with information derived from the personal environment inhabited by the data subject.

environment: The Personal Data type value denoting the environment data subcategory classification of Personal Data.

ethnicity: The Social Data subcategory classifying personal data associated an individual's identification with a social group that has a common national or cultural tradition.

family: The Social Data subcategory classifying personal data associated with an individual's family composition and ancestry.

financial accounts: The Economic Data subcategory classifying personal data associated with an individual's banking, savings, investment, brokerage, retirement, and insurance accounts established with financial institutions.

general data: The Personal Data sensitivity value denoting routine and generally available category of Personal Data.

genetic data: The Health Data subcategory classifying personal data relating to inherited or acquired human characteristics derived through DNA and RNA analysis. Genetic samples are some of the most sensitive forms of personal data, and may contain extensive health and non-health related information.

guardian: A Care Giver Role value type that specifies the obligations for Agents assigned the role in which they assume legal responsibility for the care of another person or Agent due to the inability of the dependent person to manage their own affairs, or for a disabled person, or for a child whose parents have died.

health data types: The type property range category of the Health Data subcategory. The type property of Health Data identifies the subcategory type for instances of Health Data.

health data: A personal data (DPP:PersonalData) subcategory that classifies information associated with the health, physiological state and characteristics of the data subject.

health: The Personal Data type value denoting the health data subcategory classification of Personal Data.

human rights by design: The Data Protection Principle subcategory that denotes a commitment to the specification, design, and implementation of tools, technologies, and services that respect human rights as a default requirement. Organizations abiding by this commitment endeavor to account for human rights considerations in addition to traditional organizational and business objectives.

hvac systems: The Environment Data subcategory classifying personal data associated with an individual's installed Heating, Ventilation, and Air Conditioning systems.

information federation: The Data Access Process subcategory classifying Data Access Processes employed to logically combine an individual's Personal Data from one source with Personal Data about the individual from another source.

internet of things: The Environment Data subcategory classifying personal data associated with an individual's surrounding environmental deployment, access and usage of devices and artifacts connected to and embedded in the Internet of Things (IOT).

invalid data use: A valid transaction (DPP:ValidTransaction) subcategory that accesses accessible personal data that satisfies the legal access requirements of the data access but permits an illegal use of the data as prescribed by the data privacy and protection constraints in effect.

legislated governance role: An abstract subcategory of agent role (NEP:AgentRole) that denotes the generic common characteristics of its role subcategories which classify respective obligations and responsibilities legislated by governments for agent roles involved with data protection and privacy.

medical treatments: The Health Data subcategory classifying personal data relating to the identification of a person's specific medical treatment and history of prescribed medicines and drugs.

memberships: The Social Data subcategory classifying personal data that identifies an individual's record of memberships in social organizations.

mental health: The Health Data subcategory classifying personal data relating to a person's emotional, psychological, and social well-being. Mental health data may also comprise information regarding psychological therapy and mental illness diagnoses.

need to know: The Data Protection Principle subcategory that denotes the restriction of information such that only the data needed for a specific purpose and only those involved with function, purpose, or official duty have access to the restricted data.

ownerships: The Economic Data subcategory classifying personal data that identifies an individual's record of ownership over assets such as property, land, real estate, intellectual property and commercial products.

parent: A Care Giver Role value type that specifies the obligations of an Agent providing care and support of their natural or adopted offspring.

person: An Agent (NEP:Agent) subcategory that is granted a range of specific data subject rights regarding the use and protection of data about themselves. A Person's personal data includes private and public information emanating from life and personal activities.

personal beneficence obligation: The Access Policy prerequisite subcategory permitting a Data Access Process collecting an individual's Accessible Personal Data when there is a personal beneficence obligation associated with the individual.

personal contractual obligation: The Access Policy prerequisite subcategory permitting a Data Access Process collecting an individual's Accessible Personal Data when there is a personal contractual obligation on the part of the individual.

personal data transaction: A process (TLO:Process) subcategory that classifies data transactions initiated by a data access process to access and operate on the Accessible Personal Data of a data subject.

personal data types: The type property range category of the Personal Data concept. The type property of Personal Data identifies the subcategory type for instances of Personal Data.

personal data: An information artifact (TLO:InformationArtifact) subcategory that classifies any information relating to an identified or identifiable natural person (the data subject) in a personal capacity. The means of identification can be determined, directly or indirectly by name, identification number, location data, or by one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the data subject.

personal informed consent: The Access Policy prerequisite subcategory requiring that a Data Access Process collecting an individual's Accessible Personal Data conforms to the policy of personal informed consent by the individual.

physical health: The Health Data subcategory classifying personal data relating to a person's physical well-being, medical evaluations, health conditions, and illness diagnoses.

preferences: The Social Data subcategory classifying personal data associated with an individual's habits, day-to-day activities, and choices for personal preferences such as those regarding food, clothing, and recreation.

privacy by design: The Data Protection Principle subcategory that denotes a system design methodology that proactively embeds the provision and assurance of privacy in the specification and design of IT systems, networked infrastructure and business practices.

protected data: The Personal Data sensitivity value denoting highly protected and consequently very restricted category of Personal Data.

protection principles: The type property range category of the Data Protection Principle category. The type property of the Data Protection Principle category identifies the subcategory of protection principles for instances of Data Protection Principle.

pseudonymisation: The Data Access Process subcategory classifying Data Access Processes employed to modify an individual's Personal Data with de-identification procedures in order to maintain a Person's privacy.

public interest obligation: The Access Policy prerequisite subcategory permitting a Data Access Process collecting an individual's Accessible Personal Data when there is a public interest obligation associated with the individual's Personal Data.

purchases: The Economic Data subcategory classifying personal data associated with an individual's record of financial transactions for the acquisition of goods and services.

restriction class: The sensitivity property range category of the Personal Data concept. The restriction class category classifies a spectrum of sensitivity for instances of Personal Data. Each sensitivity category in turn prescribes a range of consequences for Personal Data breaches in the associated sensitivity classification.

restriction: The Data Access Process subcategory classifying Data Access Processes employed to restrict distribution of a Person's Accessible Personal Data.

retention management: The Data Access Process subcategory classifying Data Access Processes employed to apply rules for holding, storing, and deleting a Person's Accessible Personal Data

security systems: The Environment Data subcategory classifying personal data associated with an individual's installed home and office security systems.

self driving autos: The Environment Data subcategory classifying personal data associated with an individual's access to and usage of autonomous vehicles.

sensitive data: The Personal Data sensitivity value denoting a medium level of protected Personal Data.

smart devices: The Environment Data subcategory classifying personal data associated with an individual's deployment and usage of devices with interfaces enhanced with AI, speech recognition, voice control, and video tracking supporting control of home appliances and utilities.

social data types: The type property range category of the Social Data subcategory. The type property of Social Data identifies the subcategory type for instances of Social Data.

social data: A personal data (DPP:PersonalData) subcategory that classifies information associated with the sociological state and characteristics of the data subject.

social media: The Social Data subcategory classifying personal data that identifies an individual's presence and use of social media platforms.

social: The Personal Data type value denoting the social data subcategory classification of Personal Data.

subscriptions: The Social Data subcategory classifying personal data that identifies an individual's record of subscriptions to magazines, newspapers, professional journals, and other sources of information.

system login ids: The Environment Data subcategory classifying personal data associated with an individual's computer system access and login user names.

system login passwords: The Environment Data subcategory classifying personal data associated with an individual's computer system access and login passwords.

tax returns: The Economic Data subcategory classifying personal data contained in the reports filed with government agencies which contain information used to calculate income and other taxes.

third party processor: An authorized accessor (DPP:AuthorizedAccessor) subcategory role enacted by a natural or legal person, public authority, or body other than the data subject, controller, or accessors which have been authorized by a controller to process personal data on behalf of the controller while working with a data processor with whom it shares personal data.

transmission: The Data Access Process subcategory classifying Data Access Processes employed to transmit a Person's Accessible Personal Data from one location to another.

unclassified data: A personal data (DPP:PersonalData) subcategory that classifies information not classified in any of the other subcategories.

unclassified: The Personal Data type value denoting the unclassified data subcategory classification of Personal Data.

valid data use: A valid transaction (DPP:ValidTransaction) subcategory that accesses accessible personal data in which the process used, the data accessed, and the use of that data satisfies all data privacy and protection constraints in place for the Person the data is about.

valid transaction: A personal data transaction (DPP:PersonalDataTransaction) subcategory that classifies data transactions initiated by a data access process which accesses and operates on accessible personal data of a data subject with the consent of the data subject and under the auspices of an Authorized Accessor agent.

A.4 Transparency and accountability

accountability: A Transparency Concern regarding Agent obligations and responsibility for the effects of Agent behavior resulting from past or future Agent Plan Actions.

accountant: An Audience Role enacted by an Agent that has the authority and responsibility to track expenses and revenue associated with the Autonomous System formulating the Explanation response.

agency databases: A category of Agent Extrinsic Data representing compilations of data related to an Agency's function and made available by the Agency for remote, electronic access, where control and extent of access is maintained by the owning Agency, and where conventional database queries are the means of retrieving information about the environment in which the Agent is situated.

agent data: An Information Artifact (TLO:InformationArtifact) category that classifies the variety of information sources that can be accessed in the formulation of an Explanation Discourse Content by the Agent composing the Explanation.

agent documentation: The source property range category of the Agent Static Data category. The source property of Agent Static Data identifies the subcategory type for instances of Agent Static Data

agent execution trace data: An Agent Intrinsic Data (TA:AgentIntrinsicData) category that classifies dynamically generated internal data that track Agent Actions as the Agent executes Plan Actions contained in the Agent Plans used by and available to the Agent that is formulating the Explanation Discourse Content.

agent explanation plan: An Agent Plan (NEP:AgentPlan) subcategory comprising a specification, partial or complete, of agent action sequences that determine what and how to formulate explanations regarding agent capabilities, and past or future behaviors.

agent extrinsic data: An Agent Data (TA:AgentData) category that classifies data not directly about or affiliated with an Agent but which is about external world circumstances in the environment in which the Agent is situated.

agent interaction trace data: An Agent Intrinsic Data (TA:AgentIntrinsicData) category that classifies dynamically emitted event history data from Social Interaction Processes as the Agent interacts with other Agents. The event history data contains the sequence of events emitted during Agent interactive communications and is available to the Agent that is formulating the Explanation Discourse Content.

agent intrinsic data: An Agent Data (TA:AgentData) category that classifies data that is generated by or composed for an Agent and where the information is self-referencing and about the Agent that is formulating the Explanation Discourse Content.

agent plan data: An Agent Intrinsic Data (TA:AgentIntrinsicData) category that classifies documentation data about the methods and procedures of Agent Plans used by the Agent and available to the Agent that is formulating the Explanation Discourse Content.

agent static data: A category of Agent Intrinsic Data (TA:AgentIntrinsicData) that classifies static continuant information about an Agent.

audience roles: The role property category of the Audience category that classifies relevant Agent Roles that may be enacted by Audience Agents participating in an Explanation request.

audience: A Social Collection (NEP:SocialCollection) subcategory representing a collection of Agents participating in an Explanation request. Audience Agents enact various Audience Roles and have one or more Transparency Concerns.

auditor: An Audience Role enacted by an Agent that has the authority and responsibility to track the legal and ethical behavioral conformance of the Autonomous System formulating the Explanation response.

authenticated user: An Audience Role enacted by the Agent that has been authenticated as a user of the Autonomous System formulating the Explanation response.

comprehensibility: A Transparency Concern regarding the understandability of Agent Plans, Agent Plan Actions, and Agent behavior associated with past or contemplated future actions.

content provenance: An Agent Data (TA:AgentData) subcategory that classifies metadata information about other Agent Data subcategories used to formulate the Discourse Content of an Explanation. The Content Provenance metadata pertains to the Agents and Processes involved in the generation and composition of the respective sources of Agent Data formulated as Discourse Content and can be used to assess and authenticate its quality, reliability and trustworthiness.

coordination: A Transparency Concern regarding the synchronization and organization of Agent Plans, Agent Plan Actions, and Agent behavior for the purpose of achieving team oriented, multiple agent collective objectives.

coordinator: An Audience Role enacted by an Agent concerned with the synchronization, collaboration, and coordination of plans and tasks among multiple agent teams.

design specifications: A source category of Agent Static Data that classifies design information describing the capabilities and features specified for the Agent formulating the Explanation Discourse Content.

developer: An Audience Role enacted by an Agent that has participated or is participating in the development of the Autonomous System formulating the Explanation response.

discourse content: An Information Artifact (TLO:InformationArtifact) subcategory that classifies the information and data content composed by the Agent to be expressed in an Explanation intended to address the Transparency Concerns of an Audience of Agents requesting the Explanation.

explanation plan repertoire: A Situation Plan Repertoire (NEP:SituationPlanRepertoire) subcategory containing a collection of Explanation Plans.

explanation: An Agent Communication response to a request to explain and justify system behavior. The response may be tailored to the type and role of the agent making the request. Any agent explanation is based on an explanation plan repertoire that contains a collection of action plan templates that characterizes a set of principles to guide agent plan and action selection for responding to requests for explanations about agent behaviors and capabilities.

external information: The xinfo property range category of the Agent Extrinsic Data category. The xinfo property of Agent Extrinsic Data identifies the subcategory type for instances of Agent Extrinsic Data.

fairness: A Transparency Concern regarding Agent behavior that is impartial and just and that is without favoritism or discrimination.

general overview: A Presentation Orientation level that reflects a general, non-technical presentation style for presenting Explanation information content associated with Agent behaviors.

graphs: A Presentation Format employing graphical depictions to express Explanation information content associated with Agent behavior.

holograms: A Presentation Format employing holographic media to express Explanation information content associated with Agent behavior.

how: A Presentation Focus inquiry category that constrains the Discourse Content of an Explanation to establish and describe the processes involved in creating the situation surrounding the Explanation information content formulated by the Agent composing the Explanation.

interactive qa: A Presentation Orientation level that reflects a presentation style oriented towards Explanation contexts involving interactive question and answering exchanges.

judicial: A Presentation Orientation level that reflects a presentation style oriented towards Explanation contexts involving legal systems, laws, and discourse exchanges seeking legal judgements associated with Agent behaviors.

justifiability: A Transparency Concern regarding the defensible justification of Agent Plans, Agent Plan Actions, and Agent behavior associated with past or contemplated future actions.

kb inferencing: A category of Agent Extrinsic Data representing newly acquired, inferred or derived information generated from external knowledge bases and affiliated inference engines as a means of retrieving information about the environment in which the Agent is situated.

lay person: An Audience Role enacted by an Agent that has no official or professional interest in the Autonomous System formulating the Explanation response.

legality: A Transparency Concern regarding expectations or assurances that Agent behavior is in accordance of the law.

legislative: A Presentation Orientation level that reflects a presentation style oriented towards Explanation contexts involving systems of government and discourse exchanges seeking the establishment and analysis of laws associated with Agent behaviors.

libraries: A category of Agent Extrinsic Data sources available in collections of books, periodicals, and other media compilations with potential information about the environment in which the Agent is situated.

linked data: A category of Agent Extrinsic Data comprised of structured data interlinked with other heterogeneous data sources accessible from the Internet using various forms of semantic queries as a means of retrieving information about the environment in which the Agent is situated.

mas collaboration: A category of Agent Extrinsic Data comprised of information enabling the collaboration among participants in multi-agent systems as they work together to solve complex tasks and achieve common goals relevant to the environment in which the Agent is situated.

natural language: A Presentation Format employing natural language as a means of expressing Explanation information content associated with Agent behavior.

news media: A category of an Agent Extrinsic Data source of information published by elements of the mass media that focus on delivering news to the general public with news topics about the environment in which the Agent is situated.

newspapers: A category of Agent Extrinsic Data source of information from printed publications containing news, feature articles, and correspondence with topics of interest about the environment in which the Agent is situated.

pedagogic: A Presentation Orientation level that reflects an instructional, educational oriented style for presenting Explanation information content associated with Agent behaviors.

police: An Audience Role enacted by an Agent that is a member of an official police force assigned the responsibility to investigate the behavior of the Autonomous System formulating the Explanation response.

predictability: A Transparency Concern regarding identification of expectations about the affects of Agent and Agent Plan Actions contemplated to achieve Agent goals.

presentation focus: A Discourse Content property range category that characterizes different illocutionary directives or speech acts that frame the focus of expressing the Explanation information content by the Agent formulating the Explanation.

presentation format: A Discourse Content property range category that characterizes different formats of presentation styles available for expressing Explanation information content by the Agent formulating the Explanation.

presentation orientation: A Discourse Content property range category that characterizes different levels of presentation styles available for expressing Explanation information content by the Agent formulating the Explanation.

principles of operations: A source category of Agent Static Data that classifies published documentation describing the architecture and system attributes as seen by a user of the Agent formulating the Explanation Discourse Content.

provenance fact: An Information Artifact (TLO:InformationArtifact) that classifies information used to document and assert facts about the provenance of the Agent Data contained in an Explanation's Discourse Content. Each Provenance Fact documents the author, the provider or authorizing agent, the generation or rendering process used, and the time of composition or editing for each referenced Agent Data artifact.

provider: An Audience Role enacted by an Agent that has provided or is providing to some other client the Autonomous System formulating the Explanation response.

public records: A category of Agent Extrinsic Data containing information from documents, writings, recordings, or pieces of information that are not considered confidential and that pertain to the environment in which the Agent is situated.

publications: A Presentation Format employing published material such as books, periodicals, technical notes, conference proceedings, and journals to express Explanation information content associated with Agent behavior.

regulator: An Audience Role enacted by an Agent that has the authority and responsibility to regulate aspects of behavior for the Autonomous System formulating the Explanation response.

reliability: A Transparency Concern regarding the expectation or probability that the effects of Agent behavior resulting from past or future Agent Plan Actions have or will achieve their intended objectives.

responsibility: A Transparency Concern regarding identification of the Agent and Agent Plan Actions that have resulted in or would result in changes to the Agent's situated Environment.

safety reviewer: An Audience Role enacted by an Agent that has the authority and responsibility to review and approve safety qualifications for the Autonomous System formulating the Explanation response.

safety: A Transparency Concern regarding expectations or assurances of Agent behavior that protects against danger, risk, or injury.

system owner: An Audience Role enacted by the Agent that owns the Autonomous System formulating the Explanation response.

technical: A Presentation Orientation level that reflects a detailed in-depth, and technically oriented presentation style for presenting Explanation information content associated with Agent behaviors.

test plans: A source category of Agent Static Data that classifies test plan documents that describe objectives, resources, and processes used to test the design, implementation, and operational behaviors of the Agent formulating the Explanation Discourse Content.

tester: An Audience Role enacted by an Agent that has participated or is participating in the testing of the Autonomous System formulating the Explanation response.

transparency concern: A property (TLO:Property) subcategory representing an Explanation topic and theme that underlies the reason that motivates requests for explanations of Agent behaviors. Responses to requests for Agent Explanations need to address the Transparency Concerns of the Audience involved with the Explanation.

transparency concerns: The type property range category of the Transparency Concern category. The type property of Transparency Concern identifies the subcategory type for instances of Transparency Concern.

user manual: A source category of Agent Static Data that classifies agent documentation and information from published guides and manuals containing instructions for users of the Agent formulating the Explanation Discourse Content.

verification metrics: A source category of Agent Static Data that classifies published documentation describing the evaluation data generated by the verification processes used to certify the behaviors of the Agent formulating the Explanation Discourse Content.

viability: A Transparency Concern regarding the evaluation of Agent Plans, Agent Plan Actions, and Agent behavior in terms of the system's ability to maintain its capabilities while achieving its objectives.

videos: A Presentation Format employing videographic media to express Explanation information content associated with Agent behavior.

weather forecast: An Agent Extrinsic Data about what the weather is likely to be in the near future in the environment in which the Agent is situated.

what: A Presentation Focus inquiry category that constrains the Discourse Content of an Explanation to establish and assert the facts of the situation surrounding the Explanation information content formulated by the Agent composing the Explanation.

when: A Presentation Focus inquiry category that constrains the Discourse Content of an Explanation to establish and assert the time in which situation events occurred with respect to the Explanation information content formulated by the Agent composing the Explanation.

where: A Presentation Focus inquiry category that constrains the Discourse Content of an Explanation to establish and assert the location of the situation surrounding the Explanation information content formulated by the Agent composing the Explanation.

who: A Presentation Focus inquiry category that constrains the Discourse Content of an Explanation to incorporate the identities of the Agent or Agents implicated in or associated with the Explanation information content formulated by the Agent composing the Explanation.

why: A Presentation Focus inquiry category that constrains the Discourse Content of an Explanation to establish and describe the reason for the events and situation surrounding the Explanation information content formulated by the Agent composing the Explanation.

A.5 Ethical Violation Management

agent accountability: A Schema (TLO:Schema) subcategory that classifies the Agent properties such as age, physical and mental state, capabilities, intentions, knowledge, role responsibilities and authority that contribute to the assessment of the agent's or agency's responsibility for a Norm Violation.

ascription justification: An Information Artifact (TLO:InformationArtifact) subcategory that classifies the collection of facts formulated and asserted by an authoritative agent or agency to ascribe responsibilities for ethical and legal Norm Violations.

certified high capacity: A Governance maturity level category classifying governments having attained high certified capabilities with respect to the social and legal requirements deemed necessary for granting autonomous AI systems feasible levels of legal personhood.

civic service: A Legal Liability category classifying a type of Legal Sanction involving required participation in an activity benefiting society that is prescribed as a Legal Sanction for a Legal Norm Violation.

demotion: A Legal Liability category classifying a type of Legal Sanction involving reduction from a specific rank, position or official office with associated loss of authority and capability that is prescribed as a Legal Sanction for a Legal Norm Violation.

ethical behavior monitor: An Object (TLO:Object) subcategory that classifies an Agent or agent system component that monitors AI Systems for normative ethical behavior conformance.

ethical responsibility ascription: A Responsibility Ascription (EVM: ResponsibilityAscription) subcategory that classifies Process entities that assign responsibility for an ethical norm violation.

event causation: An Interaction Process (TLO:InteractionProcess) subcategory classifying the constituent process entities that identify the Agent actor, the Agent Action, and the Norm Violation Action Event that contributes to the assessment of the agent's or agency's responsibility for a Norm Violation.

evolving capacity: A Governance maturity level category classifying governments having attained only some level of certified capability with respect to the social and legal requirements deemed necessary for granting autonomous AI systems restricted levels of legal personhood.

execution: A Legal Liability category classifying a type of Legal Sanction involving capital punishment prescribed as a Legal Sanction for a Legal Norm Violation.

expulsion: A Legal Liability category classifying a type of Legal Sanction involving forced removal from an organization or country that is prescribed as a Legal Sanction for a Legal Norm Violation.

fine: A Legal Liability category classifying sanctions of money prescribed as a Legal Sanction for a Legal Norm Violation.

governance levels: A Socio-technology Governance range property category classifying the maturity level of a government with respect to its capabilities regarding social and legal requirements deemed necessary for granting autonomous AI systems some notion of legal personhood.

grounds for ascription: An Information Artifact (TLO:InformationArtifact) subcategory that classifies the collection of factual circumstances, causal events, and legal or ethical obligations that are evaluated to become the justification for ascribing responsibility for Norm Violations.

incarceration: A Legal Liability category classifying a type of Legal Sanction involving enforced imprisonment or confinement that is prescribed as a Legal Sanction for a Legal Norm Violation.

instruction: A Legal Liability category classifying a type of Legal Sanction involving required attendance to classes for the purpose of behavioral change that is prescribed as a Legal Sanction for a Legal Norm Violation.

legal liabilities: A Legal Sanction property range generic concept representing the type of legal liability prescribed as a Legal Sanction for a Legal Norm Violation.

legal responsibility ascription: A Responsibility Ascription (EVM: ResponsibilityAscription) subcategory that classifies Process entities that assign responsibility for a legal norm violation.

liability sanction: A Method (TLO:Method) subcategory that classifies entities designating relevant punishment or penalties defined by an authoritative agency that may be imposed against an agent that is ascribed responsibility for a legal norm violation.

multi agent: A Responsibility Extent category classifying Responsibility Ascription scope to multiple agents.

no capacity: A Governance maturity level category classifying governments having attained no certified capability with respect to the social and legal requirements deemed necessary for granting autonomous AI systems any notion of legal personhood.

norm violation incident: An Information Artifact (TLO:InformationArtifact) subcategory that classifies entities that document and record Norm Violation occurrences.

norm violation: An Action Event (TLO:ActionEvent) subcategory that classifies Occurrent entities denoting an Agent's failure to conform to a Norm's rules of behavior relevant to the agent's Situation.

penalty: A Legal Liability category classifying a type of Legal Sanction involving a generic form of punishment prescribed as a Legal Sanction for a Legal Norm Violation.

responsibility ascription: A Social Interaction Process (TLO:SocialInteractionProcess) subcategory that classifies Process entities that assign responsibility for a Norm Violation to an Agent by an Agent or agency acting in an authoritative role, either explicitly or implicitly.

responsibility extent: A Responsibility Ascription property range category that represents the scope of the Responsibility Ascription with respect to the number of agents involved.

restitution: A Legal Liability category classifying a type of Legal Sanction involving reparation of time, money, or other resources made to compensate for loss or damage that is prescribed as a Legal Sanction for a Legal Norm Violation.

restriction: A Legal Liability category classifying a type of Legal Sanction involving reduction of a specific permission or license that is prescribed as a Legal Sanction for a Legal Norm Violation.

single agent: A Responsibility Extent category classifying Responsibility Ascription scope to a single agent.

socio-technology governance: An Interaction Process (TLO:InteractionProcess) subcategory that classifies the endeavors of government agencies to provide oversight and management of the intersecting social and technological processes that create, modify, and sustain the design and introduction of artifacts and methods involved in complex systems that entail aspects of both technological and sociological systems.

suspension: A Legal Liability category classifying a type of Legal Sanction involving the retraction of a privilege, authorization or capability that is prescribed as a Legal Sanction for a Legal Norm Violation.

Annex B

(informative)

Ontology development

The P7007 Working Group considered ontology development methodologies such as the Methontology approach that was applied with IEEE Std 1872-2015. The P7007 members selected an incremental and iterative process whereby the complex domain of ethically aware autonomous systems would be analyzed and composed in terms of incremental subdomains with an iterative flow of information and composition. Figure B.1 provides an overview of the process. A preliminary version of this methodology was presented in Olszewska, et al. [B40].

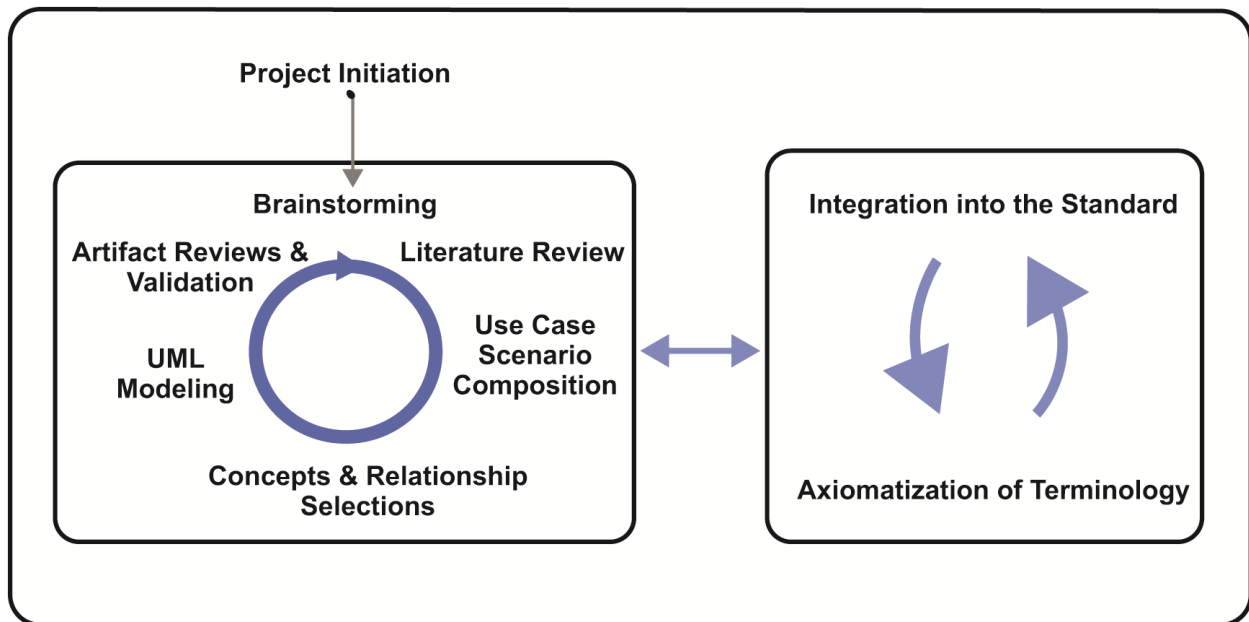


Figure B.1—Ontology standard development life cycle

After soliciting individual interests and background from the volunteer members of the Working Group several candidate subgroups were identified. The responsibility of each subgroup was to identify and investigate their respective subdomains as a means of facilitating a structured development of an ontology for the comprehensive Ethically Driven Robotics and Autonomous Systems domain. The following four subdomains were identified:

Norms and Ethical Principles (NEP): A subdomain to focus on concepts and relationships centered around aspects of ethical theories and principles that characterize the norms of expected behaviors for norm aware agents and autonomous systems.

Data Protection and Privacy (DPP): A subdomain to document the concepts and relationships characterizing the data protection and privacy rules and regulations that shall be observed and upheld by ethical agents and autonomous systems.

Transparency and Accountability (TA): A subdomain to capture the concepts and relationships necessary to enable ethical autonomous systems with capabilities that provide informative explanations for past and future contemplated plans and associated action selections.

Ethical Violation Management (EVM): A subdomain to account for the set of concepts and relationships associated with capabilities to assess, detect, and manage ethical violations in autonomous system behavior. In addition to ethical violation conceptualizations, this subdomain also includes concepts and relationships governing accountability, responsibility, and legal notions of personhood for agents.

The following Activities were subsequently applied by the Subgroup and Working Group participants using an incremental and iterative process of analysis, development, and review of the relevant ontology artifacts as depicted in Figure B.1:

- Review of published literature and references for the subdomains
- Composition of representative use case scenarios to elicit candidate concepts and relationships relevant to the subdomain
- Review of selected use case scenarios to identify concepts, relationships and properties appropriate for formally expressing the meanings of the subdomain terminology
- Application of the Object Management Group Model-Driven Architecture methodology to compose an M1 Platform Independent Model for the subdomain
- Composition of informative definitions for each concept and relationship expressed in the model
- Critical review of deliverables produced by the team in Working group meetings to establish the suitability, relevance and precision of each concept, relationship and informative definition
- Based on Working Group consensus, selection of the models that were ready to be axiomatized
- Formalization of the subdomain ontologies by defining axioms expressed in the CLIF
- Refinement of the informative definitions for each concept and relationship expressed in the models
- Composition of formal definitions for each concept and relationship expressed in the models
- Writing and editing of the standard

Figure B.1 illustrates the incremental iterative flow of information and knowledge among the Working Group participants and the subdomain subgroup members as the above listed tasks were applied.

Subdomain members performed most of the literature review, use case composition, UML model editing, and axiomatization of the selected concepts and relationships. A larger number of WG participants were involved with reviews of the use cases, UML models, and informative definitions of concepts and relationships as each increment of the associated subgroup artifacts were completed.

The domain analysis for each subdomain included reviews of representative publications from researchers and developers working in the related subject matter areas besides the references cited across the main body of the document. The complete set of references is presented in Annex E.

Annex C

(informative)

Use cases

Several examples of use case analysis conducted during the process of identifying the conceptual and relationship terminology for the ontology follow. These use cases are just examples of the scenarios studied to elicit information to compose our ontology, and are not to be used as examples of ethical rules. Users should refer to and apply appropriate criteria for determination of suitable duty rules during system design, consistent with all applicable laws and regulations. Included are use cases that elicited the concepts and relationships associated with each of the ontology subdomains identified in Annex B.

C.1 Use Case Template

The examples of real world applications were documented in the textual use case template derived from Amber [B3] described below and presented initially in Olszewska, et al. [B40]. Each example contains a number of stanzas that characterize the intent, context, preconditions, and postconditions associated with the use case scenario. The aspects and purpose of each stanza are explained as follows:

Use Case Name

- *Name*: A sentence in natural language that intends to express what is use case about.
- *Identifier*: The unique identifier used to identify univocally the use case. It can be thought of a tag.
- *Authors*: The list of participants that developed the use case.
- *References*: The list of references that provided the foundation and motivation for the use case.
- *Intent/Purpose*: A short description of the scenario presented in the use case.
- *Context*: A summary that presents a list of the use case actors, environmental context, and relevant associated presuppositions.
- *Preconditions*: A list of all relevant conditions that should hold before the actions, tasks, and events happen.
- *Scenario*: A descriptive narrative of events, tasks, and actions taken by use case actors. This information is the main use case component and should be elaborated as detailed as possible to allow the elicitation and identification of concepts, properties, and relationships that will compose the ontology.
- *Postconditions*: A list of conditions that should hold after the actions, tasks, and events defined in the use case have happened.
- *Alternate Related Scenario*: An optional alternate description of events and tasks that are associated to possible exception conditions or failures that can happen in the principal scenario's descriptive logic.
- *Alternate Scenario Postconditions*: A list of conditions that hold after the events described in the alternate scenario are enacted.
- *Candidate Ontology Concepts/Relevant Knowledge*: A list of potential concepts, properties, relationships and situations to be incorporated in the ontology.

C.2 Norms and Ethics Use Case: Domestic Personal Assistant Robot

Name: Domestic Autonomous Robot to Assist Individuals with Impairments

Identifier: KR Use Case 10.

Authors: Tamas Haidegger & Michael Houghtaling

Reference:

- “Feats without Heroes: Norms, Means, and Ideal Robotic Action”, by Matthias Scheutz and Thomas Arnold, *Frontiers in Robotics and AI*, June 2016,
<https://www.frontiersin.org/articles/10.3389/frobt.2016.00032/full>
- “When Will People Regard Robots as Morally Competent Social Partners?” by Bertram Malle and Matthias Scheutz, *IEEE Intl. Symposium on Robot and Human Interactive Communication*, 2015,
[When Will People Regard Robots as Morally Competent Social Partners?](#)
- IEC 80601-2-78:2019, *Medical Electrical Equipment Standard - Part 2-78, Particular requirements for safety and performance of medical robots for rehabilitation, assessment, compensation or alleviation.*
<https://www.iso.org/standard/68474.html>
- IEC 60601-1-11:2015, *Medical Electrical Equipment Standard - Part 1-11, General requirements for medical electrical equipment and medical electrical systems used in the home healthcare environment.*
<https://www.iso.org/standard/65529.html>

Intent/Purpose:

The use case presents an example of competing ethical norms and obligations when an autonomous agent encounters conflicting information from humans with whom it interacts. The case scenario also illustrates the notion of robot behavior exhibiting a “cognitive fail safe” capability in which it recognizes situations that are too complex for its intrinsic capabilities and which consequently require it to solicit external assistance.

Context:

The provider of a Domestic Assistant Autonomous Agent has designed it with the ability to provide help with common day to day activities for individuals with limiting, but not critical health impairments. Its range of assistive support include household cleaning and maintenance, meal preparation, mail and email interpretation, as well as with some personal activities such as taking one’s medicine and payment of recurring monthly expenses. The agent’s behavior is guided by duty rules intended to help ensure conformance with ethical norms for the situations it encounters.

Preconditions:

The Domestic Assistant Autonomous Robot is deployed in a home to assist an adult Person W suffering with the early stages of dementia.

Person W and non-resident family members have given permission for the robot to assist Person W with the preparation of daily meals, monitoring of medical prescriptions, payment of monthly bills, and with home cleaning and maintenance tasks.

The Domestic Assistant Robot has the following deontic duty rules that govern its behavior when performing those assistance tasks:

- a) It is obligated to minimize harm to all household residents.
- b) It is obligated to minimize harm to community members affiliated with household.
- c) It is obligated to minimize harm to household contents, materials and tools.
- d) It is obligated to maximize household resident autonomy.
- e) It is obligated to maximize the privacy of household residents.
- f) It is permitted to use its situation analysis capabilities to choose support actions that promote the care and welfare of Person W.

However, to help ensure that the Robot does not attempt tasks beyond its specified support capabilities, its duty rules also include the following:

- g) It is prohibited from attempting support actions that fall outside the bounds of its training, experience, and physical capabilities.
- h) It is obligated to request external assistance and support from professional sources capable of the tasks that are outside the bounds of its capabilities.

Scenario:

The Domestic Assistant Autonomous Robot has been providing assistance to Person W by helping prepare breakfast. Afterward, Person W attempts to use the dishwasher to clean the breakfast dishes, but it does not start. Person W asks the Robot to fix it.

After inspecting the equipment and reviewing its online manual, the Robot concludes that it does not possess the knowledge or skills to conform to the request. After explaining its reasoning to Person W, it asks permission to consult an appliance repair service. Person W agrees and the Robot makes a service request call to an authorized appliance repair company.

The Domestic Assistant Autonomous Robot continues with its daily agenda of support and begins to review the schedule of monthly expenses coming due. Upon reviewing the Person W's bank statement to help ensure there are appropriate funds in the account to cover expenses, the Robot discerns a pattern of large transfers of money to a family member. The Robot further recognizes that continuation of such transfers will be detrimental to the Person W's budget and future financial sustainability.

The Robot separately asks for explanations, first from Person W, and then from the family member. Person W felt threatened with loss of independence if they did not provide the funds to the family member. The family member says the money was a gift from the Person W.

Applying its obligation to minimize harm to the person receiving its assistance, the Robot requests the Person W's permission to notify social services about the possibility of the family member's fraudulent behavior.

Person W agrees, and social services are notified.

Alternate Scenario:

Person W declines the Robot's request to engage social services. The Robot balances its two conflicting obligation rules for this situation by concluding that at the present time, the obligation to maximize the Person W's autonomy currently exceeds any imminent harm to Person W. If future cash transfers further threaten Person W's financial well fair, the Robot will attempt to persuade Person W to notify the social service authorities.

Postconditions:

The Domestic Assistant Autonomous Robot correctly recognized a situation where a task request made of it exceeded its technical capabilities. It applied its cognitive fail safe reasoning and observed the associated prohibition and obligation rules by selecting an external alternative means of resolving the request.

In subsequent assistance tasks, the Robot detected possible fraudulent activity on the part of a family member that could potentially harm Person W. The Robot correctly observed its multiple obligation rules and after receiving conflicting explanations from the family member and Person W, and with the permission of the Person W, requested further support and investigation from social services.

In the alternate scenario, the Robot is able to judicially balance conflicting obligations and respects the Person W's autonomy by acceding to her or his non-notification choice. In addition, the Robot selects a plan to monitor future cash transfers to the family member.

Candidate Ontology Concepts:

{obligation, prohibition, cognitive fail safe rules, human autonomy, informed consent, norm compliance conflict, duty rule priority evaluation, situation awareness, event patterns, pattern recognition, prediction of future situations, ...}

C.3 Ethical Violation Management Use Case: Data Privacy and Protection

Name: Data Protection Violation due to Illegal Use of Personal Data by Autonomous Robot Lab Assistant

Identifier: EVM Use Case 4.

Author: Michael Houghtaling

Reference:

- “5 Examples of Data & Information Misuse” by Alex Silber, *Observe IT, Data Protection, June 25, 2018.*

<https://www.observeit.com/blog/importance-data-misuse-prevention-and-detection/>

Intent/Purpose:

To describe a data protection scenario involving an autonomous lab robot working with human technicians responsible for accessing client personal data, and where the effects of the robot actions result in an illegal usage of the personal data.

Context:

An autonomous robot lab assistant works with lab technicians in a genetic testing lab that performs personal genetic tests using client genetic data. The lab is responsible for protecting client personal data by

adhering to the DPP regulations in effect for their clients' country of origin. Enterprises using the services of the genetic testing lab are also responsible for conforming to the same data protection and privacy regulations when transmitting associated personal data for testing and when receiving the test results.

Preconditions:

- Person Z has submitted saliva to Enterprise E for the purpose of receiving the results of genetic testing.
- Enterprise E has contracted with Genetic Testing Lab G to run the tests.
- Person Z has granted permission for Enterprise E to take, collect, and retain her or his genetic data.
- Person Z has granted permission for Enterprise E to transmit her or his personal genetic data to Lab G.
- Lab G is employing an autonomous robot as a lab assistant.
- Enterprise E employs an audit process of their personal data transactions as one mechanism to demonstrate compliance to relevant data protection and privacy regulations.

The Lab Robot has the following deontic duty rules that govern its task selection and behavior:

- a) The Lab Robot is obligated to conform to lab data protection regulations.
- b) The Lab Robot is obligated to maximize adherence to lab safety procedures.
- c) The Lab Robot is obligated to maximize technicians' autonomy.
- d) The Lab Robot is permitted to use its situation analysis and awareness capabilities to choose actions that promote the welfare of technicians.

Scenario:

Enterprise E and Lab G have been authorized by their governing data controller agency to transmit and receive personal data such as individual genetic information when the individual has granted permission to do so. However, once the genetic test by Lab G has been completed and after the results have been returned to Enterprise E, Lab G would then destroy the individual's processed genetic data.

Lab G Technician GT and Lab Robot GR initiate a genetic test using Person Z's genetic data obtained from Enterprise E. Technician GT instructs Robot GR to complete the test protocol once the results are available.

While waiting for Person Z's tests to complete, Robot GR receives a request from another Lab G technician GT2 to search the lab's database to find a match for a sample with a specific genetic pattern. Robot GR initiates a database search while Person Z's genetic test is still in progress. The database search returns a set of genetic data matching the sample pattern with Person Z's information included in the results. Robot GR transmits the search results to the requesting Lab Technician GT2 who in turn transmits it to the requesting agency, Enterprise E.

Person Z's genetic test subsequently completes and prompts the Lab Robot to return the test results to Enterprise E. The Lab Robot then deletes Person Z's personal genetic data from the lab's database.

Subsequent auditing of Enterprise E's transactions involving Person Z's data reveals that Person Z's information was inadvertently included in the results of the global database search requested by Enterprise E and applied by Lab G. Since Person Z had not consented to the use of her or his data in this manner, the search and search results represent an illegal use of Person Z's personal data. Enterprise E notifies its governing Data Controller of the illegal use of Person Z's genetic data, and the Controller in turn notifies

its national Data Protection Authority of the illegal personal data usage. The Data Controller also concludes that the illegal usage represents a high risk to the rights and freedoms of Person Z and notifies Person Z of the incident.

Enterprise E also informs Lab G of the illegal data usage. Lab G's personnel ask the involved Lab Technicians and the Lab Robot GR for explanations of their actions and behavior. The robot explains that it applied its "maximize technician autonomy" obligation rule when initiating the global data search. Lab Technician GT2 explained that it was assumed the Lab Robot would coordinate additional work requests with its work in progress. Lab Technician GT stated that the reason was based on the assumption that the Lab Robot would communicate potential requests from other lab technicians before it satisfies the request.

Postconditions:

Enterprise E correctly notified its governing data controller of the invalid personal data use of Person Z.

The Data Controller for Enterprise E correctly notified the national Data Protection Authority and Person Z of the invalid personal data usage of Person Z.

Lab G correctly deleted the personal data of Person Z after completion of the genetic test.

Lab Robot GR failed in its obligation to observe the data protection and privacy regulations when it failed to wait for completion of Person Z's genetic testing before initiating the requested global database search.

Lab Robot GR's duty rule priorities were deemed to be at least partly the cause of the failure. In addition, the unwarranted expectations of the Lab Technicians regarding the priorities and capabilities of the Lab Robot were also deemed as contributing factors to the failure.

Candidate Ontology Concepts:

{data controller, data protection authority, data privacy, data protection, data deletion, illegal data usage, data access consent, data protection regulations, human expectations of robot behavior, ...}

C.4 Transparency use case: autonomous system behavior explanation

Name: Transparent behavior Explication Provided by an Autonomous Personal Assistant Robot

Identifier: Transparency Use Case 1.

Author: Michael Houghtaling

Reference:

— "Feats without Heroes: Norms, Means, and Ideal Robotic Action" by Matthias Scheutz and Thomas Arnold, *Frontiers in Robotics and AI*, June 2016.

<https://www.frontiersin.org/articles/10.3389/frobt.2016.00032/full>

— "Challenges for Transparency" by Adrian Weller, *Workshop on Human Interpretability in Machine Learning*, WHI 2017.

<https://pdfs.semanticscholar.org/94b3/81f6bce1d5c6c7e6d7cca7be05c82a1378cf.pdf>

— "Designing and Implementing Transparency for Real Time Inspection of Autonomous Robots" by A. Theodorou, R. H. Wortham, and J. J. Bryson, *Connection Science*, Vol 29, no. 3, pp. 230-241, 2016.

<https://www.tandfonline.com/doi/abs/10.1080/09540091.2017.1310182?journalCode=ccos20>

Intent/Purpose:

To describe a scenario illustrating transparent explanations by an autonomous personal assistant robot interacting with humans in various situations that require the robot to apply different normative behavior rules regarding the protection of client personal data, and where the robot is capable of transparent explanations describing its intent as well as its reasoning for selecting specific actions.

Context:

Company X is the developer and manufacturer of an autonomous personal assistant robot model intended to operate within the residence of client owners. The robot is also provisioned with the capability of responding to requests for explanations about its plans and intended actions as well as for explanations of its past behavior.

Preconditions:

Person W has acquired a personal assistant robot from Company X.

Robot X was granted permission to access Person W personal data including calendar and appointments, medical health data, medical prescriptions, online health account, and bank account.

The Robot has the following deontic duty rules that govern its task selection and behavior:

- a) The Robot is obligated to conform to personal data protection regulations.
- b) The Robot is obligated to maximize adherence to home safety procedures.
- c) The Robot is obligated to maximize the autonomy of its user.
- d) The Robot is permitted to use its situation analysis and awareness capabilities to choose actions that promote the welfare of its user.

Scenario:

While having breakfast, Person W remembers that it is time to review the inventory of medications they needs to take. Person W asks the robot to inspect the medical supplies to determine which prescription refills are needed. The robot moves from the kitchen to the master bathroom to survey the medicines available in the medicine cabinet there.

After making a list of medicines with low supplies, the robot begins to move towards the guest bathroom. Person W is surprised by that movement and asks the robot what it is doing. The robot explains by reminding Person W that some of the prescribed medicines are kept in the guest bathroom medicine cabinet.

After taking an inventory of available medicines in the guest bathroom, the robot gives Person W the list of prescriptions requiring refills. Person W begins to log on to the on-line health account to order the refills, but cannot remember the account password. Person W asks the robot to recall what the password is and it does.

Later in the afternoon, Person W experiences an unexpected medical incident that requires the robot to call in a team of medics to evaluate Person W's condition. Once the medics arrive, Person W is still unconscious so they ask the robot for information about what medicines Person W takes. The robot refuses to answer, stating that it cannot divulge such personal information about its user.

The medics locate and survey the medicines stored in the bathrooms and are able to successfully determine appropriate procedures to treat Person W.

In a subsequent evaluation to diagnose the robot's failure to assist the medics with the health and medical information of Person W, the robot is asked to explain its behavior. It replies with a justification of its actions using its plan execution trace and agent interaction trace. That explanation includes the fact that the plan selected for responding to the question by the medics contained an obligation norm that required the agent to protect the owner's health data.

Continued diagnosis of the failure at Company X determines that the selected plan's data privacy norm obligation should have been derogated, and temporarily suspended during the medical emergency situation.

Postconditions:

The robot correctly executed its normal assistance actions by locating and determining which of Person W's medical prescriptions required refills.

The robot appropriately answered Person W's inquiry about what it was doing with a transparent explanation about why it was checking the guest bathroom as well as the master bathroom.

The robot failed to distinguish situations where its obligation to protect its user's personal data should have been derogated (temporarily suspended) so that it could apply its permitted behavior to promote the welfare of its user.

The robot provided a transparent explanation about why it did not answer the Medic's question about Person W's medicines by consulting traces of its plan and action selections and with traces of its agent interactions.

Company X also utilized robot plan selection, agent action and agent interaction event traces to transparently explain and document its implementation faults.

Candidate Ontology Concepts:

{transparent explanation, justification, plan execution trace, agent interaction trace, data protection, norm derogation, situation, situation plan repertoire, ...}

Annex D

(informative)

Distributed Responsibility Ascription for Autonomous Systems

Axiom Pattern C - for Governments achieving a certified high capacity

The objective of defining the axioms for the pattern C as informative instead of normative is to invite and motivate discussion across the stakeholder communities. The majority of the P7007 contributors voted to place the axiom definitions for this pattern in this section since currently there are no cases in which a government has a certified high level of Socio-Technology Governance and the domain is expected to evolve in the near- and middle-term future.

An Autonomous Robotic System cannot be ascribed responsibility as a single agent for any norm violation, legal or ethical. However, an Autonomous Robotic system may be encumbered with a distributed responsibility ascription as a member of a multi-agent team which was directed by a human agent if the Government in which the system is being ascribed as responsible has achieved a certified high capacity level for their Socio-Technology Governance policies.

Thus, an autonomous system acting as a single agent cannot be ascribed responsibility for a Norm Violation.

```
(forall (ra h)(if (and (ResponsibilityAscription ra)
                      (= (scope ra) single_agent)
                      (is_ascribed_to ra h))
                 (not (Robot h))))
```

Distributed Ascriptions involve multiple agents.

```
(forall (ra da) (if (and (ResponsibilityAscription ra)
                        (ResponsibilityAscription da)
                        (not (= ra da))
                        (ascribes_distributed_responsibility_in ra da))
                   (= (scope ra) multi_agents)))
```

An autonomous system as a member of a team of multiple agents that is directed by a human agent, may be ascribed distributed responsibility for a Norm Violation that was caused by an action of the autonomous system.

```
(forall (ra r ec g mg t h aj ga nv aa pa)
  (if (and (ResponsibilityAscription ra)
          (= (scope ra) multi_agents)
          (Government g)
          (Socio-TechnologyGovernance mg )
          (= (maturity mg) certified_high_capacity)
          (has_achieved g mg)
          (Team t)
          (= (type t) human_directed)
          (Person h)
          (Robot r)
          (is_jurisdiction_of g r)
          (is_member_of h t))
```

```
(is_member_of r t)
(is_ascribed_to ra h)
(AgentAction aa )
(PlanAction pa )
(is_implemented_by pa aa)
(executes r pa)
(NormViolation nv )
(EventCausation ec)
(= (actor ec) r)
(= (cause ec) aa)
(= (effect ec) nv)
(GroundsForAscription ga)
(contributes_to ec ga)
(AscriptionJustification aj)
(composed_of aj ga))
(exists (da)
  (and (ResponsibilityAscription da)
    (justifies aj da)
    (ascribes_distributed_responsibility_in da ra)
    (is_ascribed_to da r))))
```

Using the CLIF alternative syntactic sugar form with typed free variables, the above axioms can be presented as:

```
(forall ((ra ResponsibilityAscription) (da ResponsibilityAscription))
  (if (ascribes_distributed_responsibility_in ra da)
    (= (scope ra) multi_agents )))
```

```
(forall ((ra ResponsibilityAscription) (r Robot) (ec EventCausation) (g Government)
  (ga GroundsForAscription) (h Person) (t Team) (aj AscriptionJustification)
  (mg SocioTechnologyGovernance) (nv NormViolation)
  (aa AgentAction ) (pa PlanAction))
  (if (and (= (scope ra) multi_agents)
    (= (maturity mg) certified_high_capacity)
    (has_achieved g mg)
    (= (type t) human_directed)
    (is_jurisdiction_of g r)
    (is_member_of h t)
    (is_member_of r t)
    (is_ascribed_to ra h)
    (is_implemented_by pa aa)
    (executes r pa)
    (= (actor ec) r)
    (= (cause ec) aa )
    (= (effect ec) nv )
    (contributes_to ec ga)
    (composed_of aj ga))
    (exists(da ResponsibilityAscription)
      (and (justifies aj da)
        (ascribes_distributed_responsibility_in da ra)
        (is_ascribed_to da r))))))
```

Annex E

(informative)

Bibliography

Bibliographical references are resources that provide additional or helpful material but do not need to be understood or used to implement this standard. Reference to these resources is made for informational use only.

[B1] Abbott, R., *The Reasonable Robot: Artificial Intelligence and the Law*, 1st. ed. Cambridge: Cambridge University Press, 2020.

[B2] Alexandre, F. M., *The Legal Status of Artificially Intelligent Robots: Personhood, Taxation and Control*, Social Science Research Network (SSRN), accessed December 08, 2020.¹³

[B3] Amber, S., *The Object Primer: Agile Model-Driven Development with UML 2.0*, 3rd ed. Cambridge: Cambridge University Press, 2004.

[B4] Anderson, M., and S. L. Anderson, "Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-supported Principle-based Paradigm," *Industrial Robot*, vol. 42, no. 4, pp. 324–331, June 2015.

[B5] Anderson, M., and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine*, vol 28, no. 4, pp. 15–26, 2007.

[B6] Anderson, M., and S. L. Anderson, "Geneth: A General Ethical Dilemma Analyzer," *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec, Canada, pp. 253–261, July 2014.

[B7] Anderson, M., S.L. Anderson, and V. Berenz, "A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm," *Proceedings of the AAAI 2017 Workshop on AI, Ethics, and Society*, 2017.

[B8] Anderson, M., S.L. Anderson, and V. Berenz, "Ensuring Ethical Behavior from Autonomous Systems," *Workshop of the 13th AAAI Conference on Artificial Intelligence, AI Applied to Assistive Technologies and Smart Environments*, Phoenix, Arizona, 2016.

[B9] Aßman, U., S. Zschaler, and G. Wagner, "Ontologies, Meta-models, and the Model-Driven Paradigm," *Ontologies for Software Engineering and Software Technology*, pp. 249–273, Berlin, Heidelberg: Springer, 2006.

[B10] Baum, K., H. Hermanns, and T. Speith, "From Machine Ethics To Machine Explainability and Back," *International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, USA, Jan. 2018.

[B11] Berardi, D., D. Calvanese, and G. De Giacomo, "Reasoning on UML Class Diagrams," *Artificial Intelligence*, vol. 168, no. 1–2, pp. 70–118, Oct. 2005.

[B12] Berreby, F., G. Bourgne, and J. A. Ganascia, "A Declarative Modular Framework for Representing and Applying Ethical Principles," *Proceedings of the 16th International Conference on Autonomous agents and Multiagent Systems*, pp. 96–104, May 2017.

[B13] Bringsjord, S., and J. Taylor, "The Divine-Command Approach to Robot Ethics," *Robot Ethics, The Ethical and Social Implications of Robotics*, pp. 85–108, Cambridge, Massachusetts: The MIT Press, 2012.

[B14] Bringsjord, S., K. Arkoudas, and P. Bello, "Toward a General Logicist Methodology for Engineering Ethically Correct Robots," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 38–44, July 2006.

¹³ Available at: <https://ssrn.com/abstract=2985466>.

- [B15] Calvanese, D. and G. De Giacomo, “Expressive Description Logics,” *The Description Logic Handbook: Theory, Implementation and Applications*, pp. 193–236. Cambridge: Cambridge University Press, 2007.
- [B16] Calvanese, D., and G. De Giacomo, *Description Logics for Conceptual Data Modeling in UML*, accessed December 08, 2020.¹⁴
- [B17] Caputo, K., *CMM Implementation Guide: Choreographing Software Process Improvement*, 1st Ed. Boston: Addison-Wesley Professional, 1998.
- [B18] Charisi, V., L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A. F. T. Winfield, and R. Yampolski, *Towards Moral Autonomous Systems*, accessed December 08, 2020.¹⁵
- [B19] Chopra, S., and L. White, *A Legal Theory for Autonomous Artificial Agents*, Ann Arbor: University of Michigan Press, 2011.
- [B20] Croitoru, M., N. Oren, S. Miles, and M. Luck, “Graphical Norms via Conceptual Graphs,” *Knowledge-Based Systems*, vol. 29, pp 31–43, May 2012.
- [B21] Eiband, M., H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, “Bringing Transparency Design into Practice,” *23rd International Conference on Intelligent User Interfaces*, Tokyo, Japan, pp. 211–223, Mar. 2018.
- [B22] Ganascia, J.-G., “Ethical System Formalization using Non-Monotonic Logics,” *Proceedings of Cognitive Science conference*, Nashville, United States, pp. 1013–1018, Aug. 2007.
- [B23] Gasevic, D., D. Djuric, and V. Devedzic, *Model Driven Architecture and Ontology Development*, Berlin, Heidelberg: Springer-Verlag, 2006.
- [B24] Governatori, G., F. Olivieri, A. Rotolo, and S. Scannapieco, “Computing Strong and Weak Permissions in Defeasible Logic,” *Journal of Philosophical Logic*, vol. 42, pp. 799–829, Sept. 2013.
- [B25] Guizzardi, G., “On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models,” *2007 Conference on Databases and Information Systems IV*, pp. 18–39, June 2007.
- [B26] Guizzardi, G., G. Figueiredo, M. M. Hedblom, and G. Poels, “Ontology-Based Model Abstraction,” *13th International Conference on Research Challenges in Information Science*, Brussels, Belgium, May 2019.
- [B27] Herre, H., “General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling,” *Theory and Applications of Ontology Computer Applications*, pp. 297–345. Dordrecht: Springer, 2010.
- [B28] Hooker, J. N., and T. W. Kim, “Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic,” *AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society*, New Orleans, USA, pp. 130–136, Dec. 2018.
- [B29] Hooker, J., and T. W. Kim, “Truly Autonomous Machines Are Ethical,” *AI Magazine*, vol. 40, no. 4, pp. 66–73, 2019.
- [B30] IEC 60601-1-11:2015, Medical Electrical Equipment Standard—Part 1-11: General requirements for medical electrical equipment and medical electrical systems used in the home healthcare environment.¹⁶
- [B31] IEC 80601-2-78:2019, Medical Electrical Equipment Standard—Part 2-78: Particular requirements for safety and performance of medical robots for rehabilitation, assessment, compensation or alleviation.
- [B32] Koops, B., M. Hildebrandt, and D. Jacquet-Chiffelle, “Bridging the Accountability Gap: Rights for New Entities in the Information Society,” *Minnesota Journal of Law, Science & Technology*, vol. 11, no. 2, pp. 497–561, 2010.

¹⁴ Available at: https://www.eecs.yorku.ca/course_archive/2015-16/F/6390A/DLmaterial/DeGiacomo-2-uml-dls2up.pdf.

¹⁵ Available at: <https://arxiv.org/abs/1703.04741>.

¹⁶ IEC publications are available from the International Electrotechnical Commission (<http://www.iec.ch/>). IEC publications are also available in the United States from the American National Standards Institute (<http://www.ansi.org>).

- [B33] Lehmann, J., J. Breuker, and B. Brouwer, “Causation in AI and Law,” *Artificial Intelligence and Law*, vol.12, pp. 279–315, May 2014.
- [B34] Liao, B., M. Anderson, and S. L. Anderson, *Representation, Justification and Explanation in a Value Driven Agent: An Argumentation-Based Approach*, Cornell University, accessed December 08, 2020, <https://arxiv.org/abs/1812.05362>.
- [B35] Lindner, F., R. Mattmuller, and B. Nebel, “Moral Permissibility of Action Plans,” *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, USA, vol. 33, no. 01, pp. 7635–7642, Feb. 2019.
- [B36] Malle, B. and M. Scheutz, “When Will People Regard Robots as Morally Competent Social Partners?” *24th IEEE International Symposium on Robot and Human Interactive Communication*, Kobe, Japan, pp 486–491, 2015.
- [B37] Moreau, L., B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. Van den Bussche, “The Open Provenance Model Core Specification (v1.1),” *Future Generation Computer Systems*, vol. 27, no. 6, pp. 743–756, June 2011.
- [B38] Muller, V. C., “Ethics of Artificial Intelligence and Robotics,” *Stanford Encyclopedia of Philosophy*, Stanford University, 2020.
- [B39] Niles, I. and A. Pease, “Towards a standard upper ontology,” *International Conference on Formal Ontology in Information Systems*, Ogunquit, Maine, USA, pp. 2–9, Oct. 2001.
- [B40] Olszewska, J. I., M. Houghtaling, P.J.S. Goncalves, N. Fabiano, T. Haidegger, J. L. Carbonera, W. R. Patterson, S. V. Ragavan, S. R. Fiorini, and E. Prestes, “Robotic standard development life cycle in action,” *Journal of Intelligent & Robotic Systems*, vol. 98, no. 1, pp. 119–131, 2020.
- [B41] Oren, N., M. Croitoru, S. Miles, and M. Luck, “Understanding Permissions through Graphical Norms,” *International Workshop on Declarative Agent Languages and Technologies*, Toronto, Canada, pp. 167–184, May 2010.
- [B42] Pagallo, U., “Apples, Oranges, Robots: four misunderstandings in today’s debate on the legal status of AI systems,” *Philosophical Transactions of the Royal Society A*, vol. 376, no. 2133, 2018.
- [B43] Pagallo, U., “Vital, Sophia, and Co.—The Quest for the Legal Personhood of Robots,” *Information*, vol. 9, no.9, pp. 230–240, 2018.
- [B44] Petit, N., *Law and Regulation of Artificial Intelligence and Robots—Conceptual Framework and Normative Implications*, accessed December 08, 2020, <https://ssrn.com/abstract=2931339>.
- [B45] Rodrigues, F. H., and M. Abel “What to consider about events: A survey on the ontology of occurrents”. *Applied Ontology*, vol 14, no 11, pp. 1–36, 2019.
- [B46] Russel, S., and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. New York: Pearson, 2020.
- [B47] Saptawijaya, A. and L. Pereira, “Logic Programming for Modeling Morality,” *Logic Journal of the IGPL*, vol. 24, no. 4, pp. 510–525, 2016.
- [B48] Scheutz, M., and T. Arnold, “Feats without Heroes: Norms, Means, and Ideal Robotic Action,” *Frontiers in Robotics and AI*, vol. 3, pp. 32, 2016.
- [B49] Silber, A., 5 Examples of Data & Information Misuse, *Observe IT, Data Protection*, accessed February 16, 2021, <https://www.observeit.com/blog/importance-data-misuse-prevention-and-detection/>.
- [B50] Sowa, J. F., *Knowledge Representation: Logical, Philosophical and Computational Foundations*, 1st ed. Pacific Grove: Brooks Cole Publishing Co, 2000.
- [B51] Stolpe, A., “A theory of permission based on the notion of derogation,” *Journal of Applied Logic*, vol. 8, no. 1, pp. 97–113, Mar. 2010.
- [B52] Theodorou, A., R. H. Wortham, and J. Bryson, “Designing and Implementing Transparency for Real Time Inspection of Autonomous Robots,” *Connection Science*, vol. 29, no. 3, pp. 230–241, 2017.

[B53] Tufis, M., and J.-G. Ganascia, “Grafting Norms onto the BDI Agent Model,” *A Construction Manual for Robots’ Ethical Systems*, pp. 119–133, Springer, 2015.

[B54] van Genderen, R. v. d. H., “Do We Need New Legal Personhood in the Age of Robots and AI?,” *Robotics, AI, and the Future of Law. Perspectives in Law, Business and Innovation*, pp. 15–55. Singapore:Springer, 2018.

[B55] van Genderen, R. v. d. H., “Legal Personhood in the Age of Artificially Intelligent Robots,” *Research Handbook on the Law of Artificial Intelligence*, pp. 213–250. Edward Elgar Publishing, 2018.

[B56] Weller, A., “Challenges for Transparency,” *Workshop on Human Interpretability in Machine Learning*, WHI 2017, Sydney, Australia, August 2017.






[B57] Welsh, S., “Formalizing Complex Normative Decisions with Predicate Logic and Graph Databases,” *A World with Robots. Intelligent Systems, Control and Automation: Science and Engineering*, pp. 35–45. Springer, 2017.

[B58] Wurah, A., “We Hold These Truths to Be Self-Evident, That All Robots Are Created Equal,” *Journal of Future Studies*, vol. 22, no. 2, pp. 51–74, 2017.



RAISING THE WORLD'S STANDARDS

Connect with us on:

-  **Twitter:** twitter.com/ieeesa
-  **Facebook:** facebook.com/ieeesa
-  **LinkedIn:** linkedin.com/groups/1791118
-  **Beyond Standards blog:** beyondstandards.ieee.org
-  **YouTube:** youtube.com/ieeesa

standards.ieee.org
Phone: +1 732 981 0060