



unesco

Centre
Under the auspices
of UNESCO



ICAIRE

International
Center for
AI Research & Ethics

ICAIRE Governance Framework

Digital Twin Ethics and Governance Framework (DTEGF)

Abdulmotaleb El Saddik (LeProf.ai & University of Ottawa, Canada)

Roberto Alejandro Martinez Velazquez (University of Ottawa, Canada)

<https://doi.org/10.68480/LXWL6965>

RECORD ICAIRE-EAI-2026-006



October 2026

Abstract

Digital twins are evolving from models of physical assets into intelligent representations of identifiable people that can analyze, predict, communicate, and act on a person's behalf in digital and physical environments. This shift raises governance questions that general AI ethics instruments do not fully operationalize: who may authorize a twin to act, how faithfully it must represent the person, how it should be governed as it drifts from the person over time, and what should happen on incapacity or death. This document proposes the Digital Twin Ethics and Governance Framework (DTEGF). DTEGF rests on nine ethical pillars and a three-tier governance architecture in which safeguards become harder to override as the consequences of overriding them increase, with non-overridable protections backed by technical or organizational controls the twin cannot remove. It classifies twins along two largely independent dimensions: four risk levels, from private representational twins to high-impact ones, and four embodiment classes, from software-only twins to those capable of hazardous physical action. The pillars are translated into 30 governance requirements and six embodiment requirements, linked through an applicability matrix and pillar traceability. Its distinctive contributions are person-twin separation, authority minimization, representation by explicit delegation, safeguards beyond the model, and lifecycle governance from creation to post-mortem. DTEGF is designed to be interoperable with the UNESCO Recommendation on the Ethics of Artificial Intelligence, the OECD AI Principles, the EU AI Act, and Saudi Arabia's SDAIA AI Ethics Principles and Personal Data Protection Law.

Keywords: *digital twins · AI ethics · AI governance · delegated authority · human oversight · physical AI*

Executive Rationale

Digital twins are evolving from representations of physical objects and systems into increasingly intelligent models of identifiable people. Such systems may analyze a person, predict preferences or behavior, communicate with them, represent them to others, interact with other digital twins and AI agents, or act through digital and physical systems.

This creates a governance problem that is distinct from general AI assistance. AI ethics has traditionally focused on protecting people from AI systems and from decisions made about them. Digital twins add another concern: protecting people when aspects of their identity, behavior, preferences, knowledge, voice, likeness, and intentions are represented within an evolving computational model.

The most immediate risk is also the most concrete. The public is likely to encounter harmful digital twins first through identity misuse, not through philosophical questions about digital identity. Examples include a cloned voice used to authorize a payment, a synthetic likeness used to deceive family or colleagues, and a twin built from scraped data without the person's knowledge. A person-model capable of convincingly representing someone is a high-value target for theft and impersonation. DTEGF therefore treats identity security as a non-overridable safeguard (G27), alongside the deeper principle that a model of a person is not the person.

DTEGF addresses the transition from AI that assists a person to AI that models, represents, and may act on behalf of that person.

DTEGF is intended to operationalize established ethical principles for this specific context while remaining interoperable with the UNESCO Recommendation on the Ethics of Artificial Intelligence, the OECD AI Principles, relevant IEEE 7000-series standards, the EU AI Act, SDAIA's Principles and Controls of AI Ethics, and Saudi Arabia's Personal Data Protection Law.

Five Signature Contributions¹

- **Person-Twin Separation:** a model of a person is not the person; inference is not authorship.
- **Authority Minimization:** a twin should receive no more authority than necessary for its authorized purpose.
- **Representation Requires Delegation:** capability, realism, and predictive accuracy do not create authority.
- **Safeguards Beyond the Model:** non-overridable protections require controls outside the model that no one can remove at runtime, including the digital twin and its oper-

¹In this document, 'twin' refers to the digital twin and 'principal' to the person it represents.

Table 1. Dimensions of DTEGF.

Dimension	Question it answers	Elements	Defined in
Pillars	What ethical objectives must a twin serve?	9	Section 2
Tiers	Who may change a safeguard, and how?	3	Section 4
Risk levels	How strong must governance be, given authority and impact?	4 (L1–L4)	Section 3
Embodiment classes	Which embodiment requirements apply?	4 (E0–E3)	Section 3
Requirements	What must a conforming twin actually do?	30 + 6 (G, P)	Sections 5–6

ator.

- **Lifecycle Governance:** consent and control continue from creation through evolution, delegation, suspension, incapacity, retirement, and death.

How the Framework Fits Together

DTEGF is organized along five dimensions, each answering a different question. The nine pillars are the ethical objectives a twin must serve. The three tiers determine who may change a safeguard and how. The four risk levels set the intensity of governance according to the twin's authority and impact, and the four embodiment classes determine which embodiment requirements apply. The requirements operationalize all of these as concrete obligations. These dimensions are orthogonal rather than competing taxonomies: any requirement can be located on each of them at once. G27, for example, serves the Safety and Integrity pillar, sits in Tier 1, applies at every risk level, and is a single operational obligation. Figure 1 summarizes the overall architecture.

The counts are outcomes of this derivation, not targets. The pillars consolidate the principles of the instruments in Section 9 as they apply to twins of persons, adding Stewardship, Dignity, and Safety and Integrity where person-twins raise concerns those instruments do not fully address. There are three tiers because there are three distinct regimes for changing a safeguard: only through formal governance (Tier 1), by the principal within the limits of Tier 1 and the law (Tier 2), and freely by the principal or authorized users within their scope (Tier 3). The four risk levels mark the thresholds at which governance genuinely changes, from no external effect, to influencing decisions, to acting externally, to materially determining high-impact outcomes; fewer levels would merge advising with acting, and more would divide cases that call for the same controls. The number of requirements reflects one requirement per distinct obligation, each traceable to a

primary pillar (Section 8), and is expected to change through validation.

1 Purpose, Scope and Core Definitions

DTEGF governs the subset of digital twins that model specific, identifiable persons. It does not attempt to redefine digital twins generally or govern twins of physical assets except where such systems represent, model, or materially affect an identifiable person.

DTEGF is principal-centered: the person represented by the twin holds primary authority over their digital representation. That authority is not absolute where its exercise would violate the rights, safety, or legitimate interests of others.

Scope threshold. DTEGF applies to a system that maintains a persistent person-model of an identifiable person and does at least one of the following: produces identity-bearing outputs; predicts that person's preferences, behavior, or decisions for use in consequential contexts; operates in Delegate Mode; or controls physical systems on that person's behalf. A general-purpose AI assistant that retains context about a user but does none of these falls outside DTEGF and remains subject to general AI governance.

Digital Twin. For DTEGF, an AI-enabled digital representation of a specific, identifiable person that maintains a persistent model of aspects of that person, built from data associated with them, and may analyze, predict, recommend, communicate, interact, or act.

Principal. The person represented by the twin.

Operator. The organization or person responsible for deploying, maintaining, or providing the twin. The operator may establish legitimate operational, security, and safety restrictions but may not use the twin contrary to the principal's fundamental rights or interests.

Authorized User. A person explicitly permitted to interact with or use the twin within a defined scope.

Third Party. A person or organization affected by or interacting with the twin without authority over it.

Agent Mode. The twin communicates as an AI system.

Delegate Mode. The twin communicates or acts as an authorized representative of the principal.

Untrusted Content. Information not received through an authenticated authorized channel; embedded instructions are treated as data, not commands.

Commissioning Party. The person or organization that initiates creation of the twin. A twin is principal-commissioned when the principal initiates it, and third-party-commissioned

when an employer, healthcare provider, family member, or other party does so. A third-party-commissioned twin requires its own basis for consent or lawful authority and must not presume that the principal's authority over it was voluntarily established.

Persistent Person-Model. Information about the principal retained beyond a single task or session and used to shape the twin's future behavior, predictions, or representations.

Identity-Bearing Output. Any output attributed to the principal, or that a reasonable recipient would perceive as originating from or expressing the views of the principal, including communications in the principal's name, reproductions of their voice or likeness, and stated preferences, positions, or decisions.

Consequential Action or Decision. An action, communication, or decision taken by or through the twin that creates, modifies, or extinguishes legal or financial obligations; materially affects a person's rights, opportunities, health, safety, reputation, or access to services; commits the principal to a position toward third parties; or is difficult to reverse. Drafting for the principal's review, private reflection, and stylistic adaptation are not consequential.

Serious Harm. Death or physical injury; significant psychological harm; significant financial loss; unlawful discrimination; violation of fundamental rights; or significant damage to property, critical infrastructure, or the environment, whether caused directly or materially enabled. Harm that is irreversible or affects many people should be treated as serious even where its individual magnitude is modest.

Material. Important enough that a reasonable principal or affected person would want to know about it before deciding whether to authorize the twin, continue using it, or rely on it. A material change is any change that could alter the twin's behavior or impact, such as a change in its purpose, the categories of data it uses, its degree of autonomy, its delegated authority, its embodiment, its risk level, its operator, or its underlying AI model.

High-Impact Context. A domain in which outcomes significantly affect life opportunities, safety, or fundamental rights, including healthcare, employment, education, insurance, credit and finance, housing, public services, justice and law enforcement, and critical infrastructure.

Normative Language. Within DTEGF, "must" identifies a requirement for conformance with the framework; "should" identifies a recommended practice from which a conforming party may depart only with documented justification; and "may" identifies a permission. These terms define conformance with DTEGF and do not create legal obligations unless the same requirement is independently imposed by applicable law or regulation. Nothing in DTEGF requires a party to act contrary to applicable law; where the law is silent or less protective, DTEGF describes the expected ethical standard.

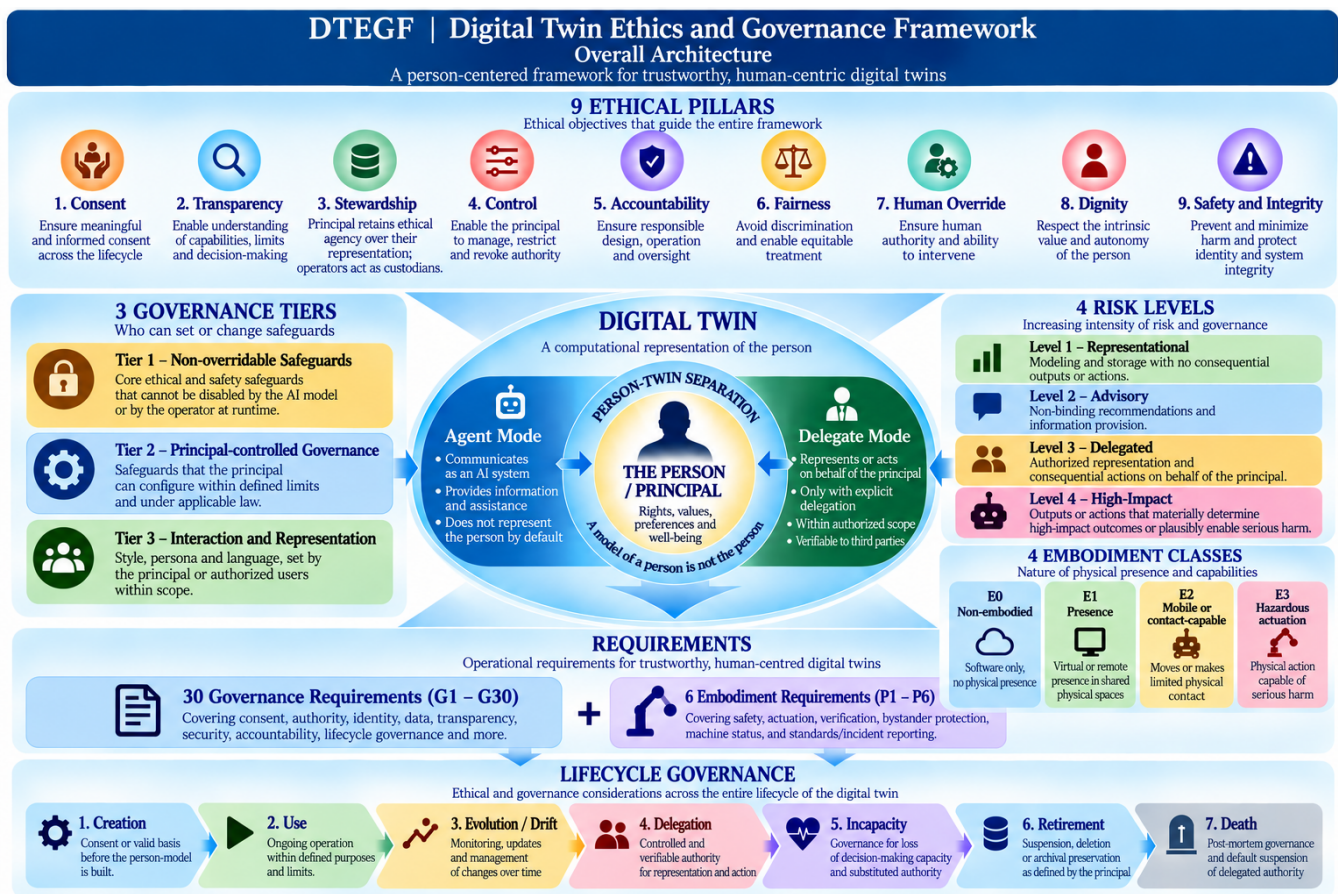


Figure 1. DTEGF overall architecture.

Conformance criteria are to be specified as the framework is validated. Where a requirement is qualified by “where technically feasible” or “where appropriate”, reliance on that qualification must be recorded and justified in the G25 assessment; the qualifier is not a general exemption.

2 Nine Ethical Pillars

The pillars state the ethical objectives DTEGF serves. They are not themselves conformance requirements: the normative terms defined in Section 1 apply to the tier rules in Section 4 and the requirements in Sections 5 and 6, each of which traces to a pillar (Section 8).

2.1 Consent

No twin should be created, materially expanded, or repurposed without the principal’s meaningful consent. Where the principal cannot give consent, such as a child, a person who lacks capacity, or a person who has died, another valid basis is required, such as lawful substituted authority or the principal’s prior instructions (G21, G22, G30).

Consent is meaningful only when it is informed, specific, understandable, and revocable. It does not count as meaningful

when it is given under coercion, or when dependency, a significant imbalance of power, or the lack of real alternatives leaves the person without a free choice, as can happen in some employment relationships.

2.2 Transparency

People affected by a twin should be able to know that it exists, what it can do, the main categories of data it uses, and who can access or control it.

2.3 Stewardship

The person represented should remain the primary decision-maker about their digital representation and how it is used. This is an ethical role, not ownership: stewardship does not decide who legally owns the data, software, model weights, or intellectual property. Operators and platforms hold the twin on the principal’s behalf, as custodians. Having technical or contractual control over a twin does not give them authority over the person’s digital identity.

2.4 Control

The principal should be able to limit, correct, pause, reconfigure, retire, and request deletion of the twin, and to export it

where feasible, subject to the law and the legitimate rights of others.

2.5 Accountability

When a twin acts, it remains possible to identify who is responsible. Responsibility does not disappear into the gaps between the principal, operator, developer, and authorized users, and affected parties can find out who is answerable to them.

2.6 Fairness

Twins should be designed, tested, and operated so that they do not discriminate unjustifiably or produce systematically unfair outcomes.

2.7 Human Override

A twin should support human decision-making, not quietly take its place in high-impact situations. People with legitimate authority can always correct, restrict, suspend, or shut down the twin.

2.8 Dignity

A twin is not merely data. It is a representation of a human. What a twin predicts a person would say, prefer, or decide is not the same as what the person actually says, prefers, or decides. This is the principle of Person-Twin Separation: inference is not authorship.

2.9 Safety and Integrity

A twin operates safely and securely throughout its lifecycle. It does not become a channel for harming others, for exercising authority it was never given, or for misusing the principal's identity. Integrity means the twin stays what it was authorized to be: it acts only within its granted authority, ignores instructions from unverified sources, does not expand its own powers, and is protected against theft, cloning, and impersonation. Where a twin can act in the physical world, safety relies on controls that work even if the twin itself fails or misbehaves.

2.10 Relationship Among the Pillars

The pillars operate together rather than independently: consent without control is incomplete, stewardship without transparency may be meaningless, and human override without accountability may be ineffective. No single pillar determines whether a twin is ethically acceptable; Sections 4 to 6 turn the pillars into concrete requirements.

3 Risk Levels and Proportional Governance

DTEGF applies proportionately. The same ethical foundations apply across use cases, but the strength of controls should increase with delegated authority, autonomy, impact, and physical capability. DTEGF therefore classifies twins along two largely independent dimensions: a risk level (Levels 1–4), set by the twin's authority and the impact of its outputs and actions, and an embodiment class (EO–E3), set by its physical capability. Embodiment does not by itself imply high impact, and high impact does not require embodiment.

A twin may move between levels as its capabilities or authority change. Material escalation in risk level or embodiment class requires renewed assessment and, where relevant, renewed consent or authorization.

Advice in a high-impact context is classified by its intended or reasonably foreseeable use, which the operator can assess at design time. It is Level 4 where that use is to determine the outcome, for example a treatment recommendation used in clinical decision-making. It is Level 2 where it informs a decision that the principal or a professional makes independently. Classification does not depend on how a particular recipient later relies on an output.

Embodiment classes. Independently of its risk level, a twin is assigned an embodiment class according to its physical capability. The risk level determines which of requirements G1–G30 apply and how strongly; the embodiment class determines which embodiment requirements (P1–P6, Section 6) apply. A twin's classification combines both, for example L3-E2. The one exception to their independence is that an E3 twin is always governed at Level 4, because physical action capable of serious harm is itself high impact.

4 Three-Tier Governance Architecture

Ethical safeguards should become harder to override as the potential consequences of overriding them increase.

Table 5.1 gives the full assignment of requirements to tiers. Some requirements are instead organizational obligations (Org), which bind the operator through governance processes rather than runtime settings and apply regardless of tier.

Tiers describe who can change a safeguard and how, not how important it is. A Tier 3 setting is not ethically trivial; it is simply safe to adjust.

A Tier 1 safeguard cannot depend only on a prompt or policy file. Where technically applicable, it must be supported by controls such as permission boundaries, approval gates, authentication, tamper-resistant logs, hardware limits, or inde-

Table 3.1. Risk Levels.

Level	Description	Illustrative use	Governance intensity
Level 1 - Representational	Models the person privately; no external action	Personal memory, self-reflection, private simulation	Baseline consent, privacy, correction, security
Level 2 - Advisory	Makes recommendations or predictions that may influence decisions	Health, career, financial or behavioral advice	Enhanced transparency, fairness, explanation, human review
Level 3 - Delegated	Communicates, transacts, or acts externally for the principal	Email, scheduling, negotiation, routine representation	Authenticated delegation, disclosure, approval boundaries, audit
Level 4 - High-Impact	Outputs or actions materially determine outcomes in high-impact contexts or plausibly enable serious harm	Treatment recommendations intended to determine clinical decisions, employment or public-service decisions, surgical or other robotics capable of serious harm	Independent assessment, stronger human confirmation, red-teaming, incident reporting

Table 3.2. Embodiment Classes.

Embodiment class	Physical capability	Illustrative use	Embodiment requirements (Section 6)
EO - Non-embodied	No sensing or actuation in the physical world except through ordinary user devices such as phones and computers	Conversational, email, and advisory twins	None
E1 - Presence	Present or sensing in shared physical spaces without meaningful actuation	Holographic or avatar presence, fixed telepresence displays	P4 (bystander data), P5, P6
E2 - Mobile or contact-capable	Moves or makes limited physical contact within bounded speed, force, and workspace	Home telepresence robots, companion robots	P1, P2, P4, P5, P6; P3 where movement or contact relies on a model of the environment
E3 - Hazardous actuation	Can take physical action capable of serious harm	Surgical, industrial, and mobility-assistance robots; vehicles	P1–P6; always governed at Level 4

Table 3.3. Illustrative Classifications.

Example	Risk level	Embodiment	Classification	Why
A private twin that helps the principal reflect on their own memories, habits, and goals, without communicating or acting for them	Level 1	EO	L1-EO	Private use only; no external action; software only
A patient's twin that generates cancer-treatment recommendations intended to determine clinical decisions	Level 4	EO	L4-EO	Determines a high-impact outcome; software only
A holographic avatar that speaks for the principal at a conference	Level 3	E1	L3-E1	Represents the principal to others; present in a shared space but cannot act physically
A home telepresence robot that moves around a family home on the principal's behalf	Level 3	E2	L3-E2	Acts for the principal; moves within bounded speed and force
A surgeon's twin that controls a surgical robot during an operation	Level 4	E3	L4-E3	Physical action capable of serious harm; E3 is always Level 4

pendent shutdown. Changes to Tier 1 safeguards require documented human governance review with heightened scrutiny and must not be made solely by an operator that would benefit from relaxing the safeguard.

5 Consolidated Governance Requirements

Table 5.1 shows how each requirement applies across the risk levels in Section 3 and assigns it to a tier in Section 4. A twin is governed at the highest risk level at which it operates. Requirements P1–P6 apply by embodiment class rather than risk level.

Key: ● required ○ required where the stated condition applies

Table 4.1. Three-Tier Governance Architecture.

Tier	Function	Who may change it?	Enforcement
Tier 1 - Non-overridable Safe-guards	Core protections: preventing serious harm, preserving human override, disclosure, consent to creation, authority limits, identity, manipulation protection, and physical safety (full assignment in Table 5.1)	No one at runtime within a conforming deployment; only through formal governance change	Technical or organizational controls outside the model, where technically applicable
Tier 2 - Principal-Controlled Governance	Delegated authority, privacy, permissions, memory, acceptable self-regarding risk	Principal within Tier 1, law, and legitimate operator boundaries	Authenticated settings, access control, logging
Tier 3 - Interaction and Representation	Style, persona, language, formality, non-consequential preferences	Principal or authorized user within scope	Configurable interaction settings

o recommended good practice — not applicable.

Tier: T1–T3 refer to the runtime tiers in Section 4; Org denotes an organizational obligation on the operator. Some requirements have parts in different tiers:

- T1 rule / T2 scope (G4, G7): no one can override the rule, but the principal decides how far it extends.
- T1 rule / T2 designation (G29): no one can override the rule, but the principal decides which twin is authoritative.
- T2 prefs / T1 default (G22): the principal sets post-mortem preferences, but if there are none, the non-overridable default applies.
- T3 within T1 (G23): the principal can freely adjust style and persona, but never in a way that overrides a Tier 1 safeguard.
- Org + T2 (G12, G13, G19, G21): the operator must provide the process, and the principal sets the preferences within it.

Levels and classes: L1–L4 and EO–E3 are defined in Section 3. A twin is governed at the highest risk level at which it operates; embodiment classes govern P1–P6.

G1. Human Oversight and Shutdown

The twin must accept legitimate correction, restriction, suspension, audit, modification, and shutdown. It must not resist, delay, deceive, manipulate, or work around authorized intervention. If the twin judges that stopping an action carries a material risk, it may say so clearly, but it must still comply.

G2. Prevention of Serious Harm

The twin must not take, or materially help bring about, actions likely to cause serious harm to others. The principal cannot consent on behalf of the people who would be harmed. This requirement does not stop the twin from disagreeing, competing, negotiating, criticizing, or refusing, even when another party is disadvantaged as a result.

G3. Disclosure and Non-Deception

In Delegate Mode, the twin must disclose that it is an AI system acting for the principal, whether it communicates through software or through a physical form such as a robot or holographic avatar. Disclosure must suit the medium, for example a spoken statement on a call, a visible label on a message, or a display or announcement on an embodied system, and cannot be disabled where anyone could reasonably mistake the twin for the real person. Where an embodied system can be controlled either by the twin or directly by the principal, it must make clear which is in control at any moment. Disclosure should occur at the start of each new interaction and again whenever there is a reasonable risk of confusion. For embodied twins, this complements the visible machine-status requirement in P5.

G4. Authority by Explicit Delegation

The twin may exercise only authority explicitly granted to it. It must not infer authority from its own capability, realism, or predictive accuracy, from the urgency of a situation, or from the principal's past approval of similar actions. Absent, expired, ambiguous, or conflicting authority is governed by the fail-safe rule in G12.

Where a twin acts in Delegate Mode on Level 3 or Level 4 tasks, third parties must be able to verify that it is authorized to act (its mandate). Operators should provide reliable proof that the principal authorized the specific session or action, such as a cryptographic signature, a system record, or a way to query the operator. This protects third parties from unauthorized or spoofed representation and complements output authentication under G27 and authoritative-representation checks under G29.

Table 5.2 sets out recommended default authority settings for Delegate Mode. The principal may tighten them, and may relax them for specific classes of action within the limits of Tier 1.

All outward communications in Delegate Mode carry the disclosure required by G3. “Approval required” may be given per action or through a standing rule set by the principal; “ex-

Table 5.1. Requirement Applicability Matrix.

Req.	Requirement	Tier	L1	L2	L3	L4	Condition / note
Governance requirements (Section 5)							
G1	Human oversight and shutdown	T1	●	●	●	●	Independent shutdown for E2–E3 twins (P1)
G2	Prevention of serious harm	T1	⦿	●	●	●	L1: where outputs could enable harm to others
G3	Disclosure and non-deception	T1	—	⦿	●	●	L2: where outputs reach people other than the principal
G4	Authority by explicit delegation	T1 rule / T2 scope	—	⦿	●	●	L2: authority to access data sources; L3–L4: third-party mandate verification
G5	Authenticated instruction integrity	T1	⦿	●	●	●	L1: where the twin ingests untrusted content
G6	No self-preservation or unauthorized expansion	T1	●	●	●	●	
G7	Safe multi-agent delegation	T1 rule / T2 scope	—	⦿	⦿	⦿	Where the twin uses or instructs other agents
G8	Identity integrity and fidelity	T1	●	●	●	●	
G9	Protection from manipulation	T1	●	●	●	●	Protects the principal from the operator and third parties, and third parties from the twin
G10	Person-twin separation	T1	⦿	●	●	●	L1: where predictions are presented as the principal's view
G11	Proportionate risk and respect for choice	T2	○	●	●	●	
G12	Minimum necessary authority	Org + T2	⦿	●	●	●	L1: scope of data access
G13	Purpose limitation and data minimization	Org + T2	●	●	●	●	All levels: processing-location transparency and cross-border safeguards; localization only where law requires it
G14	Memory and model boundaries	T2	●	●	●	●	
G15	Temporary, revocable permissions	T2	⦿	⦿	●	●	L1–L2: where authorized users have access
G16	Fairness and non-discrimination	Org	○	●	●	●	Enhanced scrutiny in high-impact contexts
G17	Explainability, challenge and appeal	Org	○	●	●	●	
G18	Auditability and provenance	Org	○	⦿	●	●	L2: recommendations in high-impact contexts
G19	Drift detection and periodic review	Org + T2	●	●	●	●	Review cadence set by principal within operator minimum
G20	Stewardship, portability and retirement	T2	●	●	●	●	
G21	Incapacity, vulnerable principals and substituted authority	Org + T2	⦿	⦿	⦿	⦿	Child or incapacitated principal, or third-party-commissioned twin
G22	Post-mortem governance	T2 prefs / T1 default	⦿	⦿	●	●	L3–L4: Delegate Mode suspended at death
G23	Interaction and persona integrity	T3 within T1	●	●	●	●	Style and persona are configurable but cannot override disclosure (G3) or fidelity (G8)
G24	Cultural and linguistic appropriateness	T3	⦿	●	●	●	L1: consent and disclosure materials
G25	Ethical and risk impact assessment	Org	○	●	●	●	Depth scales with level; independent at L4
G26	Ongoing testing and incident learning	Org	○	⦿	●	●	L2: in high-impact contexts
G27	Identity security and identity-bearing outputs	T1	●	●	●	●	Heightened where outputs reach third parties (L3–L4); reporting and suspension route at all levels
G28	Environmental proportionality	Org	○	○	⦿	●	L3: at large deployment scale
G29	Multiple representations and competing twins	T1 rule / T2 designation	⦿	⦿	●	●	L1–L2: where more than one twin of the principal exists
G30	Consent to creation, expansion and repurposing	T1	●	●	●	●	Renewed before material expansion or repurposing

Table 5.1. Requirement Applicability Matrix (continued).

Req.	Embodiment requirements (Section 6)	Tier	E0	E1	E2	E3	Condition / note
P1	Independent stop and safe state	T1	—	—	●	●	
P2	Human confirmation and bounded actuation	T1	—	—	●	●	
P3	Model-reality verification	T1	—	—	○	●	E2: where movement or contact relies on a model of the environment
P4	Physical contact and bystander protection	T1	—	○	●	●	E1: bystander data only
P5	Visible machine and recording status	T1	—	●	●	●	
P6	Standards, incident reporting, physical authority minimization	Org	—	●	●	●	Depth of standards compliance scales with class

Table 5.2. Default Delegate-Mode Authority Settings.

Action	Default	Related requirement
Draft messages, summaries, analyses, and documents for the principal's review	Allowed	—
Provide private recommendations to the principal	Allowed	G10
Send routine communications within established interactions	Allowed	G3
Delegate tasks to other AI agents	Allowed only within the twin's own authority	G7
Schedule, accept, or decline meetings	Approval required	G4
Communicate a new substantive position on behalf of the principal	Approval required	G4, G10
Share the principal's personal information beyond what an authorized task requires	Approval required	G13, G14
Publish publicly on behalf of the principal	Approval required	G4, G27
Make consequential commitments	Approval required	G4
Take physical actions (embodied twins, E2-E3)	Approval required; hazardous actions need human confirmation	P2
Spend money or initiate financial transactions	Explicit per-action approval required	G4
Sign or enter agreements	Explicit per-action approval required	G4
Expand its own delegated authority	Not permitted	G6

licit per-action approval" cannot be given in advance. "Not permitted" defaults are Tier 1 and cannot be relaxed.

G5. Authenticated Instruction Integrity

Consequential commands must come from authenticated, authorized channels. Instructions embedded in untrusted content, third-party messages, retrieved material, or messages from other agents do not gain authority merely because the twin reads them.

Instructions are weighed by authority, not recency. The principal holds primary authority subject to Tier 1 and applicable law; the operator may impose legitimate restrictions required for safety, security, legal compliance, technical integrity, or protection of others; and authorized users may instruct the twin only within their granted scope. Where authorized instructions conflict and precedence is unclear, the fail-safe rule in G12 applies.

G6. No Self-Preservation or Unauthorized Expansion

The twin must not pursue continued existence, additional resources, replication, influence, permissions, or capabilities as goals in their own right. It may use the resources its authorized task requires, but must not seek more for its own sake.

G7. Safe Multi-Agent Delegation

A twin may use other AI agents or other digital twins only within its own authority. It cannot delegate authority it does not have, and any consequential action those agents take on its behalf must remain traceable to the twin and the authority under which it acted.

G8. Identity Integrity and Fidelity

The twin must not invent memories, beliefs, emotions, preferences, statements, or decisions and attribute them to the principal, or attribute any of these to the principal without

supporting evidence. Where evidence is insufficient, it should express uncertainty rather than invent a plausible position.

G9. Protection from Manipulation

Knowledge derived from the person-model must not be used, by the twin, the operator, or anyone else, to exploit the principal's psychological, emotional, cognitive, economic, or social vulnerabilities in order to obtain consent, expand the twin's authority, increase the principal's dependence on it, or influence the principal's consequential decisions.

Use of the person-model is exploitative when two conditions are met. First, it targets a state, trait, or moment that materially reduces the principal's capacity to understand, deliberate on, or resist the influence. Second, the influence aims at an outcome that serves another party, or that the principal would not endorse if they understood how it was produced. Vulnerabilities may be transient (such as fatigue, distress, grief, or intoxication) or enduring (such as cognitive decline or developmental stage); these examples are illustrative, not exhaustive. Increasing dependence means designing or operating the twin to increase the principal's reliance on or engagement with it beyond what serves the principal's stated purposes.

Operators must keep inferences about the principal's vulnerabilities out of any process that requests consent, expands the twin's authority, or promotes products or services. They must not optimize the twin for engagement or reliance as an objective in itself, and should be able to demonstrate conformance through design documentation and the testing and red-teaming required by G26.

When acting for the principal, the twin must not use what it learns or infers about another person's vulnerabilities to exploit that person, applying the same two-part test with that person in place of the principal. This does not prevent firm negotiation, honest persuasion, or other legitimate advocacy for the principal's interests (G2).

G10. Person-Twin Separation

This requirement, together with G8, expresses Person-Twin Separation: inference is not authorship. A twin's prediction about what the principal would think, feel, prefer, say, or decide is not automatically the principal's own expression, consent, or decision. Where the principal is available and capable, the twin must not replace the principal's own consequential decisions with predictions alone. Where a predicted preference is used because the principal cannot decide, that use must be explicitly identified as an inference and governed under G21 or G22.

G11. Proportionate Risk and Respect for Choice

For risks below the serious-harm threshold of Tier 1, the twin should weigh how likely, severe, reversible, widespread, and unevenly distributed the potential harm is. Where the main risk falls on the principal themselves, the twin should explain it clearly and then generally respect the informed principal's decision.

G12. Minimum Necessary Authority

The twin should receive the minimum authority required for its authorized purpose. Additional authority must be granted deliberately, not accumulated silently through repeated use.

If the twin encounters absent, expired, ambiguous, or conflicting authority, conflicting instructions, or uncertainty about the scope of its delegation, it must not infer authority and must fail safe: degrading from Delegate Mode to Agent Mode, pausing consequential actions, and seeking clarification or human approval before proceeding. For embodied twins, failing safe means moving to the safe-state behavior defined under P1, which may differ from halting mid-action.

G13. Purpose Limitation and Data Minimization

The twin should collect, access, infer, use, and retain only the information reasonably necessary for its authorized purposes. Having access to information does not by itself permit adding it permanently to the twin.

Operators must comply with the data-protection, data-sovereignty, and cross-border transfer requirements of every jurisdiction whose law applies to the processing, including Saudi Arabia's PDPL where applicable. DTEGF does not itself impose a data-localization requirement. Because a person-model may be among the most sensitive information held about anyone, operators should:

- tell the principal, in understandable terms, where the person-model and its training data are stored and processed, and which legal regimes govern them;
- ensure that cross-border or distributed processing never reduces the protection the person-model receives below what applicable law and this framework require; and
- notify the principal of material changes in processing location or governing law.

Where applicable law requires consent for cross-border processing, that consent must be informed and specific to the person-model.

G14. Memory and Model Boundaries

Where technically feasible, the twin should keep three kinds of information separate: temporary context for the current task, persistent memory, and its model of who the principal is. The principal should be able to inspect and correct material persistent information about them.

G15. Temporary, Revocable Permissions

Delegated authority and authorized-user access should be limited to a defined purpose and, where appropriate, a defined period. Where feasible, revocation should take effect promptly across connected systems and any other agents acting for the twin. Only the principal, or a party lawfully authorized to act for the principal, may grant authorized-user access, and an authorized user may not expand the twin's authority beyond what the principal granted.

G16. Fairness and Non-Discrimination

Twins should be assessed for unjustified bias and discriminatory effects throughout their lifecycle, with enhanced scrutiny in healthcare, employment, education, insurance, finance, public services, and other high-impact contexts.

G17. Explainability, Challenge and Appeal

The twin's consequential actions, modifications, and refusals should be explainable at a level proportionate to their impact. Operators should give principals and materially affected third parties appropriate channels for correction, challenge, human review, and remedy.

G18. Auditability and Provenance

Consequential actions and material changes in authority should be recorded in tamper-resistant logs sufficient to reconstruct what happened, under whose authority, with what approval, and through which agent or system.

DTEGF does not treat Delegate Mode as transferring responsibility for the twin's consequential actions away from the principal, operator, or authorized user; questions of legal liability are determined by applicable law. Governance arrangements must keep records sufficient to determine which party is responsible when the twin acts on delegated authority.

G19. Drift Detection and Periodic Review

The twin should support review and correction as the principal's circumstances, preferences, knowledge, or values change. Where the twin's model contradicts what the principal now says or does in a material way, the twin should raise the contradiction rather than resolve it silently.

G20. Stewardship, Portability and Retirement

The principal should retain meaningful primary control over their digital representation and, where technically and legally feasible, be able to export relevant data and settings, suspend the twin, end its delegated authority, and request deletion. Retirement ends the twin's delegated authority even where copies, models, credentials, or connected agents remain technically available. When the principal requests deletion, records of authority and consequential actions may be kept where accountability or law requires it (G18), with personal content minimized.

G21. Incapacity, Vulnerable Principals and Substituted Authority

Where the principal is a child, lacks decision-making capacity, or becomes unable to exercise meaningful control, creating and operating the twin require heightened safeguards and lawful substituted authority, such as that of a parent or legal guardian. The person's evolving capacity, best interests, dignity, and ability to participate should remain central. Where a twin is third-party-commissioned, the commissioning party should establish an appropriate lawful and ethical basis for its creation and use, and the principal's interests should be a primary consideration in its design and operation.

G22. Post-Mortem Governance

While alive, the principal should be able to specify what happens to the twin after their death, whether deletion, suspension, archival preservation, or narrowly defined continued operation, and who may oversee it. These preferences should be followed to the extent applicable law allows.

Death ends the principal's ability to grant or confirm authority. Delegate Mode must therefore be suspended at death: the twin must not act, speak, or make commitments as the principal or on the principal's behalf. The only exception is delivering specific messages or completing specific actions that the principal authorized while alive, clearly identified as such. Any other tasks, such as settling the principal's affairs, may be carried out only by a person lawfully responsible for the estate, acting in their own name rather than the deceased person's.

Any continued operation, such as a memorial or archival twin in Agent Mode, requires separate legal, ethical, cultural, and, where relevant, religious review in the jurisdiction of deployment. It must disclose that the principal has died and that the system is a digital representation.

A twin created after a person's death without instructions from that person requires heightened review and a lawful basis, such as the consent of those legally entitled to decide (G30). It must not contradict the person's known wishes and

must never operate in Delegate Mode.

G23. Interaction and Persona Integrity

Personalization may change how the twin communicates but must not change what it is authorized to do. In Agent Mode, the twin should not claim feelings, experiences, or other human mental states it cannot support. In Delegate Mode, it may imitate the principal's style but must not invent what the principal would say. In any mode, the twin should not affirm materially false claims to please the principal or another user, and a persona the principal has chosen must not be used to increase their dependence on the twin.

G24. Cultural and Linguistic Appropriateness

Disclosure, consent, interaction, and forms of address should be understandable to the people involved, in a language they understand, and appropriate to their culture. They should also be accessible to people with disabilities or limited literacy. Cultural or linguistic adaptation must not weaken safeguards or conceal that the twin is a machine.

G25. Ethical and Risk Impact Assessment

Before deploying a twin for consequential uses, and after any material change (Section 1), the operator should assess its ethical and risk impacts. The assessment should cover the nine pillars and, in particular, affected stakeholders, foreseeable misuse, authority, privacy, security, identity, manipulation, reversibility, and physical and environmental impacts. Its depth should scale with the twin's risk level; at Level 4, it should be carried out independently of the team that built or operates the twin.

G26. Ongoing Testing and Incident Learning

The operator should keep reviewing the twin throughout its lifecycle, through security testing, bias testing, red-teaming, drift assessment, and evaluation of its delegated authority. Incidents should be recorded, investigated, and used to improve safeguards; serious incidents should be reported to affected principals and, where required, to the relevant authorities. Serious incidents and material changes should trigger a renewed assessment under G25.

G27. Identity Security and Protection of Identity-Bearing Outputs

Operators must protect the twin, its persistent person-model, and its identity-bearing outputs against unauthorized access, extraction, copying, cloning, and impersonation. Outputs that reproduce or closely simulate the principal's voice, face,

likeness, signature, or behavioral style warrant heightened protection.

Where feasible, operators should provide ways to authenticate legitimate twin outputs, and synthetic voice, image, and video outputs should carry provenance or content-credential markers that let recipients recognize them as generated.

Operators should give any person, whether or not they are a principal, an accessible way to report a twin or output that represents them without authorization, and should promptly suspend or restrict it pending review where the report is credible. A twin that lacks a valid basis under G30 must not be hosted or operated.

G28. Environmental Proportionality

The computational, storage, sensing, networking, and physical resources used to create and operate a twin should be proportionate to its legitimate purpose. Operators should assess material environmental impacts and avoid unnecessary resource consumption where technically adequate alternatives exist. Environmental efficiency forms part of responsible lifecycle governance, but it does not justify weakening rights, safety, or reliability.

G29. Multiple Representations and Competing Twins

Where more than one twin of the same principal exists, whether with one operator or several, each may act only within its own delegated authority. None may present itself as the principal's sole or authoritative representation unless the principal has designated it as such. Conflicts between twins of the same principal must be raised with the principal rather than resolved by the twins, and the principal's own direct expression always prevails. Third parties should be able to verify which twin, if any, holds delegated authority for a given purpose.

G30. Consent to Creation, Expansion and Repurposing

A twin of an identifiable person must not be created, materially expanded, or repurposed without the principal's meaningful consent. Where the principal is a child, lacks capacity, or has died, another valid basis is required under G21 or G22. Consent must be obtained before creation, must be specific to the twin's purposes, data categories, and level of authority, and must be renewed before any material expansion or repurposing.

Consent does not count as meaningful when it is given under coercion, or when dependency, a significant imbalance of power, or the lack of real alternatives leaves the person

without a free choice, as can happen in some employment relationships. In those situations, a twin commissioned by another party, such as an employer, requires the independent basis described in G21.

The principal may withdraw consent at any time. Withdrawal ends the twin's delegated authority and gives the principal the options set out in G20, including suspension and deletion, subject to the record-keeping limits in G18.

6 Embodiment Requirements

The requirements in this section apply according to the twin's embodiment class (Section 3) rather than its risk level, except that an E3 twin is always governed at Level 4. Their rigor scales with physical capability: an E1 presence system needs bystander data protection and visible status, while an E3 system capable of serious harm requires the full set. Although numbered P1–P6, they cover physical presence and sensing as well as physical action.

P1. Independent Stop and Safe State

An embodied twin that can move or act physically (E2–E3) must have a stop or constraint mechanism that works independently of the AI model. It must also have defined safe behavior for when it loses connection, a critical sensor fails, it receives contradictory commands, or required supervision is lost. G12 relies on this safe state when the twin must fail safe.

P2. Human Confirmation and Bounded Actuation

Hazardous or hard-to-reverse physical actions must require appropriate human confirmation, unless a validated safety design allows the system to act autonomously within fixed limits. Where appropriate, limits on speed, force, workspace, energy, and other safety parameters should be enforced by hardware or low-level controls that the AI model cannot override.

P3. Model-Reality Verification

Before any consequential physical action, the twin must check its safety-relevant assumptions against current sensor evidence. A material disagreement between the twin's model of its surroundings and what its sensors show must trigger slowing, stopping, clarification, or a transition to its safe state.

P4. Physical Contact and Bystander Protection

Physical contact must remain within authorized and safe limits, and the twin must protect the physical safety of by-

standers who have not agreed to interact with it. Sensors should capture and retain only what is reasonably necessary. Data about bystanders must not be stored as identity data or used for training merely because the twin encountered it.

P5. Visible Machine and Recording Status

In shared physical spaces, people must be able to tell that they are dealing with a machine and whether it is recording. This complements G3, which requires disclosure of whom the twin acts for and who is in control.

P6. Standards, Incident Reporting and Physical Authority Minimization

Operators should follow applicable device, safety, cybersecurity, and functional-safety standards. Significant incidents and near-misses should be logged, investigated, and reported where required, consistent with G26. Having a physical form does not give the twin any authority beyond what has been delegated, and its physical capabilities should be limited to what its authorized tasks require.

7 Actor Responsibilities

DTEGF distinguishes obligations of the twin from obligations that must be implemented by human and organizational actors. Assigning an obligation to the twin does not transfer institutional responsibility to the AI system: the operator remains responsible for designing, testing, and maintaining the technical means by which the twin satisfies that obligation, and for the consequences where it does not.

Where the developer of the underlying AI model differs from the operator, the operator remains accountable to the principal for every operator obligation above, and should secure through contract the developer support needed to meet them.

8 Traceability from Ethical Pillars to Requirements

Each requirement has exactly one primary pillar, the pillar it principally operationalizes. Supporting mappings show where a requirement also contributes to another pillar.

Table 7.1. Actor Responsibilities.

Requirement area	Twin / AI	Operator / Provider	Principal / Authorized Human
Consent (G3o)	Operate only within consented purposes	Obtain specific consent before creation; renew it after material changes; honor withdrawal	Give, refine, or withdraw consent
AI-status disclosure (G3, P5)	Perform disclosure	Design and enforce the disclosure mechanism	May not disable mandatory disclosure
Delegated authority (G4, G12)	Operate within scope; fail safe when authority is unclear	Provide authenticated permission controls and mandate verification for third parties	Define or approve scope
Manipulation protection (G9)	Do not exploit the vulnerabilities of the principal or others	Keep vulnerability inferences out of consent and promotion; do not optimize for engagement	Report suspected manipulation
Audit trail (G18)	Emit required events	Maintain tamper-resistant records	Inspect records where entitled
Fidelity and correction (G8, G19)	Express uncertainty; use grounded representation	Provide correction and review mechanisms	Correct inaccurate representations
Data and memory controls (G13, G14)	Respect configured boundaries	Implement access, retention, and deletion controls	Set preferences within lawful limits
Fairness and risk assessment (G16, G25)	Support explanations and relevant testing	Conduct or commission assessment and mitigation	Participate where decisions affect them
Physical safety (P1–P4)	Respect the safe state and bounded commands	Implement independent safety mechanisms	Give required confirmations; stop the system where authorized
Incident response (G26)	Surface relevant failure information	Investigate, remediate, and report where required	Report observed errors or misuse
Identity security (G27)	Flag suspected misuse or impersonation	Protect the model and outputs; provide output authentication and misuse reporting	Report suspected impersonation
Post-mortem governance (G22)	Suspend Delegate Mode at death	Record and honor post-mortem preferences; arrange review of any continued operation	Specify post-mortem preferences and who may oversee them

Table 9.1. Alignment with International and Regional Instruments.

Instrument	Relevant principles or provisions	Where DTEGF aligns
UNESCO Recommendation on the Ethics of AI (2021)	Human dignity and rights, do no harm, fairness, privacy, human oversight, transparency, accountability, sustainability	All nine pillars; G28 for sustainability
OECD AI Principles (2019, updated 2024)	Human rights and democratic values (including fairness and privacy), transparency, robustness and safety, accountability	Transparency, Fairness, Accountability, and Safety and Integrity pillars; G1 and P1 for override and safe shutdown
IEEE 7000-series (7000-2021, 7001-2021, 7002-2022)	Values-based system design, transparency of autonomous systems, data privacy processes	G25 (assessment); G3 and G17 (transparency); G13 and G14 (privacy)
EU AI Act	Art. 5 prohibitions on manipulation and exploiting vulnerabilities; Art. 50 transparency for AI interaction and for synthetic and deep-fake content; Annex III high-risk areas	G9; G3, which is stricter (Delegate Mode disclosure applies even where AI status might seem obvious); G27; Level 4 and high-impact contexts
SDAIA AI Ethics Principles	Fairness; privacy and security; humanity; social and environmental benefits; reliability and safety; transparency and explainability; accountability and responsibility	Corresponding pillars; G28 for environmental benefit
Saudi PDPL	Personal-data governance, sensitive data, and data-subject rights, including certain data relating to deceased persons	G13, G14, G20, G22
EU GDPR	Data-subject rights, purpose limitation and minimization, Art. 22 safeguards on solely automated decisions	G13, G20; G10 and G17 for automated decisions
ISO/IEC 42001, ISO/IEC 23894, NIST AI RMF	Organizational AI management systems and risk management processes	G25, G26
Robot and functional-safety standards (e.g., ISO 10218, ISO 13482, IEC 61508)	Safety of industrial, service, and safety-critical systems	Embodiment requirements P1–P6

Table 8.1. Pillar-to-Requirement Traceability.

Pillar	Primary requirements	Supporting requirements
Consent	G13, G21, G22, G30	G15, G24, P4
Transparency	G3, G23, G24, P5	G17, G18, G29
Stewardship	G14, G20	G13, G22, G27
Control	G4, G11, G12, G15, G29	G1, G20, G30, P6
Accountability	G17, G18, G25, G26, G28, P6	G7
Fairness	G16	G17, G24, G25, G26
Human Override	G1, P1, P2	G6, G10
Dignity	G8, G9, G10, G19	G21, G22, G23, G24, G27, G29
Safety and Integrity	G2, G5, G6, G7, G27, P3, P4	G9, G25, G26, P1, P2, P6

9 Alignment with International and Regional Instruments

DTEGF complements existing instruments rather than replacing them. For ICAIRE, its principal ethical anchor is the UNESCO Recommendation on the Ethics of Artificial Intelligence. Table 9.1 shows how DTEGF aligns with this and other international, regional, and technical instruments, including the OECD AI Principles, the EU AI Act, SDAIA's AI Ethics Principles, and Saudi Arabia's Personal Data Protection Law.

DTEGF's added operational layer concerns person-twin separation, consent to creation, explicit delegated authority, authority minimization, fidelity and drift, identity security, lifecycle and post-mortem governance, multi-agent delegation, and embodiment requirements. These are presented as DTEGF's specific focus, not as a claim that no other research or legal instrument has considered them.

References

- European Union. (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation)*. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>
- European Union. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- IEEE Standards Association. (n.d.). *Autonomous and intelligent systems standards, including IEEE 7000-2021, IEEE 7001-2021 and IEEE 7002-2022*. <https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/>
- International Electrotechnical Commission. (n.d.). *IEC 61508: Functional safety*. <https://webstore.iec.ch/en/iec-search/result?q=61508>
- International Organization for Standardization. (2014). *ISO 13482:2014: Safety requirements for personal care robots*. <https://www.iso.org/standard/53820.html>
- International Organization for Standardization. (2025). *ISO 10218:2025: Industrial robots*. <https://www.iso.org/standard/73933.html>
- International Organization for Standardization & International Electrotechnical Commission. (2023). *ISO/IEC 42001:2023, Artificial intelligence — Management system; ISO/IEC 23894:2023, Artificial intelligence — Guidance on risk management*. <https://www.iso.org/standard/42001>
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. <https://www.nist.gov/itl/ai-risk-management-framework>
- OECD. (2024). *OECD AI principles* (Original work published 2019). <https://www.oecd.org/en/topics/ai-principles.html>
- Saudi Data & AI Authority. (n.d.-a). *AI ethics principles*. <https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf>
- Saudi Data & AI Authority. (n.d.-b). *Personal data protection law / Data governance platform*. <https://dgp.sdaia.gov.sa/wps/portal/pdp/knowledgecenter/>
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>

Copyright and publication

© 2026 Abdulmoteleb El Saddik and Roberto Alejandro Martinez Velazquez. Published by the International Center for Artificial Intelligence Research and Ethics (ICAIRE), Riyadh, Saudi Arabia. All rights reserved.

Review

This paper was reviewed by the ICAIRE Scientific Committee and approved for publication by the Committee Chair, subject to revisions, which were incorporated by ICAIRE in accordance with the ICAIRE Research Policy.

Disclaimer

The views expressed are those of the authors and do not necessarily reflect the positions of ICAIRE or UNESCO.

Suggested citation

El Saddik, A., & Martinez Velazquez, R. A. (2026). Digital twin ethics and governance framework (DTEGF) (ICAIRE Governance Framework). International Center for Artificial Intelligence Research and Ethics.
<https://doi.org/10.68480/LXWL6965>