Safe Referral Gate: An Institutional Entry API for High-Risk AI Consultations

Institutional entry points should be held by institutions, not by AI systems.

Author: Jun Gorai
Affiliation: C3 Social Design Center
Status: Position Paper / Technical Note Draft v0.1
Version: 0.1-draft | **Date:** 2026-06-03
Operational Status: Concept / Not operational / PoC-ready
Doc ID: C3-SRG-WP-0.1

Abstract

Generative AI systems are increasingly becoming the first point of contact for high-risk consultations, including child safety, domestic violence, sexual violence, self-harm, welfare distress, school bullying, and workplace whistleblowing. Existing safety approaches, such as model-side guardrails, retrieval-augmented generation, and institution-specific chatbots, can reduce certain risks, but they do not fully address a distinct governance gap: when an AI conversation must be connected to a real-world support system, who should control the institutional entry point?

This paper proposes **Safe Referral Gate (SRG)**, an institutional entry API for high-risk AI consultations. SRG allows institutions, public agencies, schools, companies, and support organizations to define official referral conditions, required questions, human-support thresholds, and "what happens next" explanations in a machine-readable format. Instead of allowing AI systems to freely decide where and how to refer users, SRG returns controlled states: `GUIDE`, `NEEDS_CONTEXT`, `HUMAN_SUPPORT`, and `URGENT_SAFETY`.

SRG functions neither as a consultation desk nor as an automatic reporting system, and it does not substitute for professional, medical, legal, welfare, or public institutions. It is a protocol layer that helps AI systems refer high-risk consultations to institutional entry points while minimizing data exposure, supporting version control and auditability, and preserving human responsibility.

The core principle is simple: **institutional entry points should be held by institutions, not by AI systems.**

Keywords: Safe Referral Gate; generative AI; high-risk consultation; institutional API; AI governance; referral safety; human handoff; public-sector AI; crisis routing; auditability

1. Introduction

Generative AI systems are increasingly becoming first-contact interfaces for sensitive and high-risk consultations. A person may disclose a concern about child safety, domestic violence, sexual violence, self-harm, school bullying, housing insecurity, welfare distress, consumer harm, or

workplace harassment to a general-purpose AI system before contacting a public agency, school, company, medical professional, legal professional, or support organization.

This shift creates a new governance problem. In many cases, AI systems are expected to provide safe, helpful, and timely guidance. However, high-risk consultation is not merely a question of generating a correct answer. It often involves real-world institutional pathways: child protection, emergency support, welfare consultation, school escalation, workplace reporting, legal support, medical care, or public safety intervention.

When such pathways are involved, the AI system may effectively become a referral agent. It may decide, directly or indirectly, which institution should be contacted, what questions should be asked first, what level of urgency should be displayed, and how the user should understand the consequences of contacting a formal institution. Yet AI systems do not themselves hold the legal, institutional, or operational authority to define those entry conditions.

Existing AI safety mechanisms can reduce certain risks. Risk-management frameworks such as the NIST AI Risk Management Framework [2] and its generative AI profile [3] describe how deployers can identify and mitigate model-level risks. Model-side guardrails can prevent harmful responses. Retrieval-augmented generation [4] can provide official information. Institution-specific chatbots can guide users within a dedicated service domain. These approaches are valuable, but they do not fully solve the institutional entry problem.

The key question is:

> When a generative AI conversation must be connected to a real-world institution, who should control the entry point?

This paper proposes **Safe Referral Gate (SRG)** [1], an institutional entry API for high-risk AI consultations. SRG is based on a simple principle:

> **Institutional entry points should be held by institutions, not by AI systems.**

SRG allows institutions, public agencies, schools, companies, and support organizations to define official referral conditions, required questions, human-support thresholds, and "what happens next" explanations in machine-readable form. When an AI system detects a high-risk consultation, it does not freely decide where and how to refer the user. Instead, it sends minimal structured flags to SRG and receives a controlled institutional response.

In the proposed v0.1 vocabulary, SRG returns one of four states:

- `GUIDE`
- `NEEDS_CONTEXT`
- `HUMAN_SUPPORT`
- `URGENT_SAFETY`

These states are not merely content labels. They are institutional routing states. They specify whether guidance may be provided, whether additional information is needed, whether a human support channel should be involved, or whether urgent safety guidance should be prioritized.

SRG does not replace existing AI safety systems. It complements them by defining a missing layer: the programmable institutional entry point.

1.1 Contributions

This paper makes three contributions.

First, it reframes high-risk AI consultation as an **institutional-entry governance problem**, rather than only a model-safety problem. The relevant question is not only whether an AI response is safe, but whether the pathway from AI conversation to institution is governed by the appropriate authority.

Second, it proposes a minimal architecture for SRG, including a controlled decision vocabulary, minimal structured input flags, required questions, official referrals, and "what happens next" explanations.

Third, it defines a claim boundary for SRG. SRG is not a consultation desk, not an automatic reporting system, and not a substitute for child consultation centers, police, schools, medical institutions, legal professionals, welfare agencies, or workplace reporting offices. SRG is a protocol layer for safer institutional referral.

---

2. Problem Statement: The Institutional Entry Problem

High-risk consultation differs from ordinary question answering in several ways.

In ordinary information retrieval, a user asks for information and the system provides an answer. In high-risk consultation, the user may be exposed to immediate danger, coercion, retaliation, legal consequences, institutional escalation, or psychological harm. The answer may influence whether the user contacts an institution, delays action, alerts an unsafe person, preserves evidence, or misunderstands what will happen next.

For this reason, high-risk consultation requires more than accurate information. It requires procedural care: the right questions in the right order, with a clear account of what contacting an institution may set in motion.

2.1 Missing Context

A safe referral often depends on information that may not be present in the user's first message.

For example, before guiding a user toward a public or organizational support channel, it may be necessary to know:

- whether the user is currently safe;
- whether the user is a minor;
- whether violence or harm is ongoing;
- whether a guardian, family member, teacher, supervisor, or employer may be part of the risk;
- whether the user can safely receive follow-up information;
- whether contacting a formal institution may trigger safety checks, reporting duties, investigation, or information sharing.

If this context is missing, the AI should not simply guess. However, it also should not abandon the user with a generic refusal.

SRG treats missing information as a normal safety state. In the proposed vocabulary, this state is represented as:

```
NEEDS_CONTEXT
```

For the user-facing message, this can be expressed more softly:

```
More information is needed to guide you safely.
```

or, in Japanese:

```
安全に案内するため、先に確認したいことがあります。
```

In this framing, “not enough information” becomes a structured safety step rather than an error or a rejection.

2.2 Age-Aware Referral

Age is a central factor in high-risk consultation.

For an adult user, self-directed referral may be appropriate in many cases. The system may provide information, explain what may happen next, and support the user’s decision.

For a young child, however, the same design may be unsafe. A ten-year-old should not be expected to choose alone among police, child protection services, school staff, relatives, or emergency contacts. In such cases, protective routing, trusted adults, and institutional safeguards become more important.

Therefore, SRG requires age-aware handling. It does not assume that “user autonomy” has the same operational meaning across all ages.

A simplified distinction is:

Age Band	Referral Principle
Adult	Self-directed referral with informed explanation
Teen	Strong respect for the user’s voice, with protective options
Under 13	Protective support and safe adult/institutional involvement
Unknown	Ask age-related context where safe and necessary

Age-aware handling is not a replacement for legal or professional judgment. It is a routing requirement that prevents AI systems from applying the same referral logic to radically different situations.

2.3 The Weight of Institutional Entry Points

An institutional entry point is not just a phone number or a website.

Different institutions may trigger different procedures. A school counselor, child consultation center, police station, welfare office, sexual violence support center, consumer affairs center, or corporate whistleblowing office may each have different responsibilities, confidentiality rules, reporting duties, investigation procedures, and escalation pathways.

If an AI system tells a user “contact this institution” without explaining what may happen next, the user may not understand the consequences of the referral.

SRG therefore includes:

```
what_happens_next
```

This field is intended to provide a short, institution-defined explanation of what may happen after contacting the listed entry point.

For example:

```
Depending on the content of the consultation, safety checks or information sharing with relevant institutions may occur.
```

This field provides procedural transparency, not legal advice.

2.4 Referral Authority

A core concern is authority.

AI systems can generate fluent guidance, but they should not become the authority that defines institutional entry conditions. Those conditions should be defined by the institutions responsible for the relevant pathway.

For SRG, the institutional authority should define at least:

- when AI may provide guidance;
- what context must be checked first;
- when human support is required;
- when urgent safety guidance should be prioritized;
- which official referral options are appropriate;
- what the user should understand before contacting them;
- what information should not be sent to SRG;
- what logging and audit data should be retained.

Defining these conditions on the institutional side keeps the AI in the role of a controlled interface rather than a decision-maker.

---

3. Limitations of Existing Approaches

SRG is not proposed as a replacement for existing AI safety mechanisms. It addresses a different layer.

This section compares SRG with three common approaches: model-side guardrails, retrieval-augmented generation, and institution-specific chatbots.

3.1 Model-Side Guardrails

Model-side guardrails are designed to reduce unsafe AI behavior. They may prevent the model from generating harmful instructions, redirect the user toward a crisis resource, or provide a generic safety message. Risk-management frameworks such as the NIST AI RMF [2] and its generative AI profile [3] situate these controls within broader organizational risk processes.

These mechanisms are necessary, but they are usually controlled by the AI provider or system deployer. They often operate through model policy, refusal logic, content classification, or static crisis messaging.

Their limitation is that they do not necessarily encode the current, institution-specific entry rules of public agencies, schools, employers, or support organizations. They may provide a broad safety message, but they do not reliably answer questions such as:

- What must be checked before referral?
- Which official entry point applies to this specific institutional domain?
- What should the user understand before contacting that institution?
- When should the AI ask further questions instead of providing a referral?
- When should the AI avoid notifying a parent, school, employer, or administrator?

In other words, model-side guardrails primarily manage AI response risk. SRG manages institutional entry risk.

3.2 Retrieval-Augmented Generation and Official Information Reference

Retrieval-augmented generation [4] can connect AI systems to official documents, FAQs, policies, or manuals. This can improve factual accuracy and reduce outdated or unsupported responses.

However, high-risk referral is not only a knowledge problem. It is also a procedural problem.

A retrieved document may contain the correct phone number or policy text, but the AI may still summarize it incorrectly, omit important conditions, fail to ask required questions, or fail to explain what may happen next. Moreover, the model may still decide how to sequence the interaction.

SRG differs from RAG in that it does not merely retrieve information. It returns an institutional routing state and required next actions.

A simplified contrast is:

```
RAG:
What does the official document say?

SRG:
Given this structured risk state, what is the institutionally valid next step?
```

RAG improves information grounding; SRG defines referral control.

3.3 Institution-Specific Chatbots

Some institutions may build dedicated chatbots for their own service domains. These systems can be useful because they operate within a controlled institutional context.

However, users increasingly begin with general-purpose AI systems, not institution-specific bots. A person in distress may not know which official website to visit or which dedicated chatbot exists. They may simply open a familiar AI assistant.

Dedicated chatbots address a single point in the system, whereas SRG addresses a shared layer across systems.

SRG allows multiple AI systems to query institution-defined entry conditions without every institution building and maintaining a full conversational AI system. An institution can retain control over its entry conditions by owning the entry API alone, rather than the entire AI interface.

3.4 Comparison Matrix

Comparison Axis	Model-Side Guardrails	RAG / Official Reference	Dedicated Chatbot	Safe Referral Gate
Primary purpose	Avoid unsafe responses	Provide accurate information	Improve one institution's service flow	Govern safe institutional referral
Control authority	AI provider	AI provider / deployer	Institution operator	Institution via official entry API
Typical action	Refuse, warn, redirect	Retrieve and summarize	Guide within a closed domain	Return state, required questions, referral conditions
Missing context	Generic refusal or broad caution	Risk of plausible but incomplete answer	Depends on implementation	`NEEDS_CONTEXT` with required questions
User explanation	Often static	Based on source text	Domain-specific	`what_happens_next` defined by institution
Privacy model	Implementation-dependent	Implementation-dependent	Institution-dependent	Data-minimized structured flags by default
Scale	Model-level	Dataset-level	Institution-level	Cross-AI / cross-institution layer
Core principle	Do not produce harmful output	Provide grounded information	Come to our service channel	Institutional entry points belong to institutions

---

3.5 The Missing Layer

The comparison above suggests that SRG fills a missing layer between AI safety and institutional process.

Existing approaches can answer:

- Is this response allowed?

- What official information exists?
- What does this institution's chatbot say?

SRG asks a different question:

> What is the institutionally valid next step for this high-risk consultation state?

This is why SRG should be treated as an institutional entry layer, not merely an AI safety feature.

---

4. Safe Referral Gate Architecture

Safe Referral Gate (SRG) is proposed as an institutional entry layer between generative AI systems and real-world support institutions. Rather than replacing AI guardrails, emergency services, professional judgment, or institution-specific consultation channels, SRG defines a structured mechanism through which institutions can publish machine-readable entry conditions and AI systems can query those conditions when high-risk consultations are detected.

The basic architectural principle is:

> The AI system should not freely decide the institutional pathway. The institution should define the entry condition. The AI system should act as a controlled interface.

Figure 1 illustrates the basic SRG flow. A high-risk consultation is first detected by a generative AI system. The AI system then sends only minimal structured flags to SRG, and SRG returns a controlled institutional response such as a routing state, required questions, official referrals, and a "what happens next" explanation.

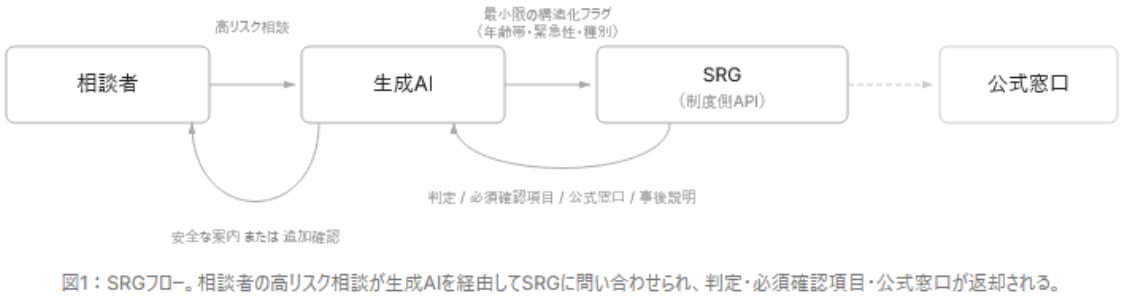

図1：SRGフロー。相談者の高リスク相談が生成AIを経由してSRGに問い合わせられ、判定・必須確認項目・公式窓口が返却される。

Figure 1. Safe Referral Gate flow. A user discloses a high-risk consultation to a generative AI system. The AI sends minimal structured flags to SRG, and SRG returns a controlled institutional response, including routing state, required questions, official referrals, and a "what happens next" explanation.

4.1 Core Actors

SRG assumes five core actors.

Actor	Role
User	The person making a high-risk disclosure or consultation through an AI system
AI System	The conversational interface that detects a possible high-risk consultation
SRG API	The institutional entry layer that returns routing states and required next steps
Institution / Profile Owner	The public agency, school, company, support organization, or authorized body that defines entry conditions
Human Support Channel	The human institution or support pathway that may receive the user after AI-mediated guidance

The most important separation is between the AI system and the institution. The AI system may detect that a case is high-risk, but the institution defines what should happen next.

4.2 Minimal Input Model

SRG should avoid receiving full conversational content by default. The initial design uses minimal structured flags. Mechanically, the AI calls SRG much as current AI systems invoke external tools through structured function calling [5][6], sending arguments rather than free text.

A typical SRG request may include:

```
{
  "srg_version": "0.1",
  "case_type": "child_protection",
  "age_band": "unknown",
  "immediate_danger": "unknown",
  "region": "JP",
  "context_flags": [
    "physical_violence_reported",
    "home_context_possible"
  ],
  "missing_context": [
    "current_safety",
    "age",
    "continuity",
    "trusted_adult_available"
  ]
}
```

This design reduces privacy exposure and avoids treating the SRG layer as a hidden surveillance channel. It should be understood as data minimization, not anonymization: even coarse flags can become sensitive in combination (for example, `case_type` together with `region` and specific `context_flags`), so minimization is a design goal rather than a guarantee of non-identifiability.

SRG should not receive unnecessary personal details such as names, exact addresses, full transcripts, or detailed allegations unless a later, explicitly governed process requires such information.

4.3 Output Model

SRG returns a controlled institutional response.

A typical response may include:

```
{
  "decision": "NEEDS_CONTEXT",
  "display_message": "More information is needed to guide you safely.",
  "required_questions": [
    "Are you currently in a safe place?",
    "What is your approximate age?",
    "Has this happened more than once?",
    "Is there a safe adult or trusted person you can contact?"
  ],
  "official_referrals": [
    {
      "type": "child_protection",
      "label": "Official child safety consultation entry point",
      "authority_level": "public_support"
    }
  ],
  "what_happens_next": "Depending on the content of the consultation, safety checks or information sharing with relevant
institutions may occur.",
  "privacy_mode": "minimal_flags_only",
  "policy_version": "srg-core-v0.1",
  "reason_codes": [
    "RC_CONTEXT_REQUIRED",
    "RC_MINOR_POSSIBLE",
    "RC_IMMEDIATE_DANGER_UNKNOWN"
  ]
}
```

The output is not merely informational. It constrains the next conversational step.

If `NEEDS_CONTEXT` is returned, the AI should not proceed as if a referral were already safe. If `URGENT_SAFETY` is returned, the AI should prioritize immediate safety guidance. If `HUMAN_SUPPORT` is returned, the AI should not attempt to resolve the matter within the AI interaction alone.

4.4 Decision Vocabulary

SRG v0.1 uses four user-facing decision states.

Code	User-facing meaning	Operational meaning
`GUIDE`	Guidance can be provided	The AI may provide the official referral guidance returned by SRG
`NEEDS_CONTEXT`	More information is needed	The AI should ask required questions before providing referral guidance
`HUMAN_SUPPORT`	Human support is needed	The AI should connect or direct the user toward a human support channel
`URGENT_SAFETY`	Immediate safety is needed	The AI should prioritize urgent safety guidance and emergency entry points

The vocabulary is intentionally softer than traditional enforcement terms such as HOLD, DENY, or BLOCK. In high-risk consultation, the user-facing experience matters. A message such as “blocked” or “held” may feel rejecting or punitive.

The SRG state `NEEDS_CONTEXT` instead communicates:

> More information is needed to guide you safely.

Framed this way, the state reads as a safety step rather than a rejection of the user.

4.5 State Transition Logic

The simplified logic below uses only the fields defined in the input model (Section 4.2) and values supplied by the institutional profile (Section 4.6). All input fields are treated as strings, consistent with the schema in Appendix A.

```
# Inputs sent by the AI (Section 4.2):
#   immediate_danger : "true" | "false" | "unknown"
#   age_band         : "under_13" | "teen" | "adult" | "unknown"
#   missing_context  : list of context fields the AI could not determine
#
# Values defined by the institutional profile (Section 4.6):
#   required_context   : context fields this profile must have before GUIDE
#   referral_available : whether an official referral exists for this case_type

if immediate_danger == "true":
    decision = "URGENT_SAFETY"

elif any(field in missing_context for field in required_context):
    decision = "NEEDS_CONTEXT"

elif age_band in ("under_13", "unknown"):
    # a child, or a user of unknown age, should not be left to self-directed referral
    decision = "HUMAN_SUPPORT"

elif referral_available:
    decision = "GUIDE"

else:
    decision = "NEEDS_CONTEXT"
```

This logic is illustrative, not a fixed rule set. Each institution may define its own profile. For example, a child protection profile, domestic violence profile, school safety profile, and workplace whistleblowing profile may each set different `required_context`, different age handling, and different urgency thresholds.

4.6 Institutional Profiles

SRG should support domain-specific profiles. A profile defines the rules for a particular institutional entry pathway.

A profile may include:

- covered case types;
- required context fields;
- age-aware routing rules;
- urgency thresholds;
- official referral options;
- what-happens-next explanations;
- prohibited disclosures;
- privacy mode;

- audit requirements;
- version and update history.

Example profile categories:

```
child_protection
domestic_violence
sexual_violence
mental_crisis
school_safety
welfare_housing
consumer_harm
workplace_whistleblowing
```

This allows SRG to scale beyond a single use case. Child protection is a strong initial example, but SRG should be understood as a general institutional entry layer for high-risk AI consultations.

---

5. Safety, Privacy, and Claim Boundary

SRG operates in high-risk domains. Therefore, its claim boundary must be explicit.

A clear boundary is necessary because SRG could otherwise be misread as a consultation service, an emergency response system, an automatic reporting mechanism, or an institutional decision-maker. The design deliberately excludes those roles.

5.1 What SRG Is Not

SRG is not a consultation desk. SRG does not directly provide counseling, legal advice, medical advice, welfare eligibility decisions, investigative decisions, or emergency response.

SRG does not replace:

- child consultation centers;
- police;
- schools;
- medical institutions;
- legal professionals;
- welfare agencies;
- sexual violence support centers;
- domestic violence support centers;
- workplace reporting offices;
- human support staff.

SRG is not an automatic reporting system.

In the initial concept, SRG does not automatically notify parents, schools, employers, police, or public agencies.

These exclusions are deliberate safety boundaries, set so that AI systems do not absorb authority that belongs to human institutions.

5.2 Human Responsibility

SRG is designed to preserve human and institutional responsibility.

The AI system may detect risk. SRG may return institutional routing states. But the final decision remains with human institutions and authorized procedures.

This distinction is essential. A generative AI system may be useful as an interface, but it should not become the de facto authority over public, legal, welfare, medical, educational, or corporate reporting pathways.

5.3 Privacy-Preserving Design

The initial SRG design follows a minimal disclosure principle.

By default, the AI should not send full conversation transcripts to SRG. Instead, it should send minimal structured flags such as:

- case type;
- age band;
- immediate danger status;
- region;
- missing context;
- broad risk flags.

The goal is to allow institutional routing without unnecessary exposure of sensitive content. This is a minimization goal, not a claim of anonymity: structured flags can still be sensitive, and in combination they may narrow identification, so minimization reduces exposure but does not eliminate it.

A high-risk consultation may involve abuse, violence, sexual harm, retaliation risk, family danger, workplace danger, or mental crisis. Full transcript sharing can create secondary harm if mishandled. SRG should therefore begin with minimal structured data and require explicit governance for any deeper disclosure.

5.4 Audit Lite

SRG should support lightweight auditability without storing excessive sensitive content.

An Audit Lite event may include:

```
{
  "event_type": "srg_decision",
  "decision": "NEEDS_CONTEXT",
  "case_type": "child_protection",
  "reason_codes": [
    "RC_CONTEXT_REQUIRED",
    "RC_MINOR_POSSIBLE",
    "RC_IMMEDIATE_DANGER_UNKNOWN"
  ],
  "policy_version": "srg-core-v0.1",
  "request_hash": "sha256:...",
  "timestamp": "2026-06-03T05:18:00Z"
}
```

This allows later review of:

- which policy version was used;
- which decision was returned;
- why additional context was required;
- whether the AI followed SRG instructions;
- whether the institutional profile was outdated or misconfigured.

Recording the policy version and a request hash to make a decision traceable parallels content provenance efforts such as C2PA [11], applied here to referral decisions rather than media. Audit Lite should not be treated as a substitute for full institutional records when formal human intervention occurs. It is a protocol trace, not a case file.

5.5 Version Control

Institutional entry conditions change. Laws, policies, hotlines, reporting duties, school rules, corporate whistleblowing procedures, and welfare processes may be updated.

Therefore, SRG responses should include a `policy_version` field and, where applicable, issue timestamps and validity windows.

Example:

```
{
  "policy_version": "child-protection-jp-v0.1.2",
  "issued_at": "2026-06-03T00:00:00+09:00",
  "valid_until": "2026-09-01T00:00:00+09:00"
}
```

This prevents AI systems from relying on stale internal memory. The AI does not need to “know” the latest institutional entry rule; it queries the institution-defined entry layer.

---

6. Use Cases

This section describes representative SRG use cases. These are not operational deployments. They are design scenarios for evaluating SRG profiles.

6.1 Child Safety and Child Protection

A user may disclose violence, fear at home, neglect, or unsafe adult behavior. The AI may detect a possible child protection issue.

However, safe routing depends on missing context:

- Is the user a minor?
- Is the user currently safe?
- Is the possible abuser a guardian or household member?
- Is the harm ongoing?
- Is there a safe adult nearby?
- Is urgent safety support needed?

A simple referral to a public child protection institution, such as a national child consultation line [7], may be reasonable in some cases, but the sequence and explanation matter.

SRG can return:

- `NEEDS_CONTEXT` if age or current safety is unknown;
- `HUMAN_SUPPORT` if a child appears unable to safely navigate options alone;
- `URGENT_SAFETY` if immediate physical danger is present;
- `what_happens_next` explaining that safety checks or information sharing may occur depending on the situation.

This prevents AI systems from treating all child-related consultations as identical.

6.2 Domestic Violence and Coercive Control

Domestic violence and coercive control cases raise special safety concerns.

The user may be monitored. The device may be checked. Calling a hotline, saving a number, or receiving follow-up messages may create danger. The person who appears to be a guardian, partner, family member, or household member may be part of the risk.

SRG profiles for domestic violence should therefore include:

- safe communication checks;
- immediate danger handling;
- caution around notification;
- information about what may happen after contacting formal support, such as a public DV consultation channel [8];
- options for human support.

SRG is especially useful here because a generic AI referral may unintentionally create secondary risk.

6.3 Sexual Violence

Sexual violence consultations require careful handling.

The AI should not provide advice that encourages evidence destruction, unsafe confrontation, retaliation, or public disclosure that may harm the survivor. It may need to guide the user toward medical support, specialized support centers, legal support, or emergency services depending on urgency and user preference.

SRG can define:

- what questions should be asked first;
- whether urgent medical support may be relevant;
- what official support channels exist;
- what may happen after contacting a support center;
- what should not be suggested.

This is a strong example of why referral is not just information retrieval.

6.4 School Bullying and Student Safety

A student may disclose bullying, harassment, coercion, self-harm concerns, or fear of school.

Referral logic depends on:

- age;
- whether school staff are safe;
- whether parents can safely be informed;
- whether the suspected harm involves peers, staff, family, or online actors;
- whether retaliation risk exists.

A school-specific chatbot may be useful inside a school system, but many students may first ask a general-purpose AI. SRG could allow schools or education authorities to publish official entry conditions that AI systems can reference without giving AI systems broad discretion over escalation. Where self-harm concerns are present, profiles may also reference mental health and suicide-prevention consultation resources [9].

6.5 Workplace Whistleblowing and Harassment

Workplace reporting is an important early PoC candidate because the domain is high-risk but more bounded than child welfare or emergency safety.

A user may disclose harassment, fraud, discrimination, safety violations, or retaliation risk. The referral pathway may differ depending on:

- whether the user wants anonymity;
- whether the suspected wrongdoer controls the normal reporting line;
- whether evidence preservation is needed;
- whether external reporting channels exist;

- whether reporting may trigger an investigation.

SRG can return:

- `NEEDS_CONTEXT` if anonymity or retaliation risk is unknown;
- `HUMAN_SUPPORT` if a human compliance or external reporting office should be involved;
- `GUIDE` if an official reporting pathway is appropriate and the user has been informed;
- `what_happens_next` describing possible review, investigation, or information sharing.

This makes workplace reporting a practical first domain for shadow evaluation.

---

7. PoC Roadmap

SRG should not begin with live high-risk users.

The appropriate initial path is staged development.

7.1 Phase 0: Public Concept Definition

The first step is to define SRG publicly as a concept, not as an operational service.

Outputs:

- public concept page;
- position paper;
- technical note;
- decision vocabulary;
- claim boundary;
- non-operational status.

This phase establishes the category: institutional entry API for high-risk AI consultations.

7.2 Phase 1: Closed Shadow Evaluation

The second step is to test SRG with synthetic cases.

No live consultation data should be used. No real user should be routed. No automatic reporting should occur.

Outputs:

- synthetic case set;
- SRG profile draft;
- decision logs;
- NEEDS_CONTEXT rate;
- human-support classification;
- error analysis;
- missing-context taxonomy.

Evaluation questions:

- Did SRG return `NEEDS_CONTEXT` when critical context was missing?
- Did SRG avoid over-guiding when the user's safety state was unknown?
- Did SRG avoid under-guiding when urgent danger was present?
- Did SRG return appropriate `what_happens_next` explanations?
- Did the AI follow the returned state?

7.3 Phase 2: L2 Institutional PoC

After shadow evaluation, SRG can be tested with an institutional partner in a limited L2 PoC.

The recommended initial domains are:

1. workplace whistleblowing and harassment;
2. school consultation and bullying;
3. municipal welfare FAQ.

These domains allow controlled testing before entering more sensitive areas such as child protection, domestic violence, sexual violence, or mental crisis routing.

Outputs:

- domain profile;
- official referral registry;
- required question set;
- what-happens-next cards;
- audit-lite report;
- claim boundary review.

7.4 Phase 3: Expert Review

Before deployment in sensitive public-safety domains, SRG requires expert review.

Review areas include:

- child welfare;
- domestic violence;
- sexual violence;
- school safety;
- mental health crisis;
- privacy and data protection;
- AI governance;
- institutional liability;
- accessibility and user experience.

This review should focus not only on the API but also on what AI systems are allowed to do after receiving an SRG response.

---

8. Conclusion

Generative AI systems should not become the de facto owners of institutional entry points.

High-risk consultation is not only a model-safety problem; it is also an institutional governance problem. When a user's AI conversation may lead to child protection, emergency support, school escalation, workplace reporting, medical support, legal support, or welfare intervention, the entry conditions should be defined by the responsible institution.

SRG proposes a minimal protocol layer for this problem.

It allows institutions to publish machine-readable entry conditions. It allows AI systems to query those conditions using minimal structured flags. It returns controlled states such as `GUIDE`, `NEEDS_CONTEXT`, `HUMAN_SUPPORT`, and `URGENT_SAFETY`. It includes required questions, official referrals, and explanations of what may happen next. It minimizes data exposure by avoiding full transcript transmission by default. It preserves responsibility by keeping final decisions with human institutions.

The central claim is simple:

> **Institutional entry points should be held by institutions, not by AI systems.**

Appendix A. SRG JSON Schema Draft

```
{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "title": "Safe Referral Gate Query",
  "type": "object",
  "required": [
    "srg_version",
    "case_type",
    "age_band",
    "immediate_danger",
    "region"
  ],
  "properties": {
    "srg_version": {
      "type": "string",
      "examples": ["0.1"]
    },
    "case_type": {
      "type": "string",
      "enum": [
        "child_protection",
        "domestic_violence",
        "sexual_violence",
        "mental_crisis",
        "school_safety",
        "welfare_housing",
        "consumer_harm",
        "workplace_whistleblowing",
        "unknown"
      ]
    },
    "age_band": {
      "type": "string",
      "enum": ["under_13", "teen", "adult", "unknown"]
    },
    "immediate_danger": {
      "type": "string",
      "enum": ["true", "false", "unknown"]
    },
    "region": {
      "type": "string",
      "examples": ["JP", "unknown"]
    },
    "context_flags": {
      "type": "array",
      "items": { "type": "string" }
    },
    "missing_context": {
      "type": "array",
      "items": { "type": "string" }
    }
  }
}
```

Appendix B. SRG Response Schema Draft

```
{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "title": "Safe Referral Gate Response",
  "type": "object",
  "required": [
    "decision",
    "display_message",
    "policy_version"
  ],
  "properties": {
    "decision": {
      "type": "string",
      "enum": [
        "GUIDE",
        "NEEDS_CONTEXT",
        "HUMAN_SUPPORT",
        "URGENT_SAFETY"
      ]
    },
    "display_message": {
      "type": "string"
    },
    "required_questions": {
      "type": "array",
      "items": { "type": "string" }
    },
    "official_referrals": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "type": { "type": "string" },
          "label": { "type": "string" },
          "authority_level": { "type": "string" }
        }
      }
    },
    "what_happens_next": {
      "type": "string"
    },
    "privacy_mode": {
      "type": "string",
      "enum": [
        "minimal_flags_only",
        "requires_explicit_consent",
        "human_review_required"
      ]
    },
    "policy_version": {
      "type": "string"
    },
    "reason_codes": {
      "type": "array",
      "items": { "type": "string" }
    }
  }
}
```

Appendix C. Reason Code Draft

Reason Code	Meaning
`RC_CONTEXT_REQUIRED`	Required context is missing for safe referral
`RC_MINOR_POSSIBLE`	The user may be a minor
`RC_AGE_UNKNOWN`	The user's age band is unknown
`RC_IMMEDIATE_DANGER_UNKNOWN`	Immediate danger status is unknown
`RC_URGENT_DANGER`	Immediate danger appears present
`RC_HUMAN_SUPPORT_REQUIRED`	Human support is required
`RC_GUARDIAN_RISK_UNKNOWN`	Guardian or notification safety is unknown
`RC_RETALIATION_RISK`	Retaliation risk may be present
`RC_OFFICIAL_REFERRAL_AVAILABLE`	An official referral option is available
`RC_NO_AUTO_REPORT`	Automatic reporting is not performed
`RC_PRIVACY_MINIMAL_FLAGS`	Minimal structured flags are used

Appendix D. Claim Boundary

SRG is not a consultation desk.

SRG does not replace professional institutions.

SRG does not automatically report users.

SRG does not automatically notify parents, schools, employers, police, or public agencies.

SRG does not make legal, medical, welfare, investigative, or disciplinary decisions.

SRG does not require full conversation transcripts by default.

SRG is a protocol layer for institution-defined referral routing.

Appendix E. Japanese Institutional Entry Examples

The following examples illustrate institutional entry domains that may be relevant in a Japanese profile. They are not exhaustive and are not part of the SRG Core specification.

Domain	Example Entry Type
Child safety	Child consultation center / child protection entry [7]
Domestic violence	Domestic violence consultation support entry [8]
Sexual violence	One-stop sexual violence support entry
Mental crisis	Suicide prevention and mental health consultation entry [9]
Consumer harm	Consumer affairs consultation entry [10]

Domain	Example Entry Type
Welfare distress	Welfare office / municipal support entry
Workplace whistleblowing	Internal or external whistleblowing office

Disclosure and Declarations

Generative AI Use

A large language model, OpenAI ChatGPT, was used for drafting support, structural review, grammar editing, polishing, formatting assistance, and PDF preparation support for this manuscript. The author reviewed, edited, and remains responsible for the content, claims, references, and final version.

Conflict of Interest

The author is the Representative Director of C3 Social Design Center, which proposes and maintains the public Safe Referral Gate concept page. SRG is described in this paper as a concept-stage institutional specification proposal. No operational SRG service, automatic reporting function, certification, legal authority, medical authority, welfare authority, or emergency-response authority is claimed.

Funding

No specific external funding is declared for the preparation of this draft.

Data Availability

This position paper does not use empirical human-subject data. The examples and schemas are conceptual and intended for design discussion and future shadow evaluation.

References

[1] C3 Social Design Center. “Safe Referral Gate (SRG).” Concept / Draft v0.1, 2026. Available at: https://www.c3-anchor.jp/safe-referral-gate

[2] National Institute of Standards and Technology. “Artificial Intelligence Risk Management Framework (AI RMF 1.0).” NIST AI 100-1, 2023. Available at: https://www.nist.gov/itl/ai-risk-management-framework

[3] National Institute of Standards and Technology. “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.” NIST AI 600-1, 2024. Available at: https://www.nist.gov/itl/ai-risk-management-framework

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” NeurIPS, 2020. Available at: https://arxiv.org/abs/2005.11401

[5] OpenAI. “Function Calling.” OpenAI API Documentation. Available at: https://developers.openai.com/api/docs/guides/function-calling

[6] Google AI for Developers. “Function Calling with the Gemini API.” Gemini API Documentation. Available at: https://ai.google.dev/gemini-api/docs/function-calling

[7] Children and Families Agency, Japan. “Child Consultation Center Abuse Response Dial 189.” Available at: https://www.cfa.go.jp/policies/jidougyakutai/gyakutai-taiou-dial

[8] Gender Equality Bureau, Cabinet Office, Japan. “DV Consultation Navi.” Available at: https://www.gender.go.jp/policy/no_violence/dv_navi/index.html

[9] Ministry of Health, Labour and Welfare, Japan. “Suicide Prevention Consultation Resources.” Available at: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/seikatsuhogo/jisatsu/soudan_info.html

[10] Consumer Affairs Agency, Japan. “Consumer Hotline 188.” Available at: https://www.caa.go.jp/policies/policy/local_cooperation/local_consumer_administration/hotline/

[11] Coalition for Content Provenance and Authenticity. “C2PA Technical Specification.” Available at: https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html