

Connor Leahy — DEDDA analysis specimen

ACCM Deep Ethics Project

Contents

1 Connor Leahy	1
1.1 Who Connor Leahy is	1
1.2 Why he is related to the ACCM Deep Ethics Project	2
1.3 Full source object	2
1.3.1 Published chapter map	2
1.4 The triad event behind this page	3
1.5 DEDDA pass 1 — preserve claim type	3
1.6 DEDDA pass 2 — what overlaps with the Project	5
1.6.1 Incentives without invented inner motive	5
1.6.2 Responsibility rises with power and correction capacity	5
1.6.3 Black boxes in the model and in governance	5
1.6.4 Psychological interaction effects	5
1.6.5 Regulation as a process object	5
1.7 DEDDA pass 3 — what remains under-developed	5
1.7.1 Omnidirectional scrutiny	5
1.7.2 The limits of transferred analogies	5
1.7.3 Narrow does not automatically mean safe	6
1.7.4 Control needs an object and an authorized controller	6
1.7.5 The four personal rules are broader than the demonstrated cases	6
1.8 DEDDA pass 4 — who regulates the regulators?	6
1.8.1 Definition power	7
1.8.2 Distributed power can still hide one cluster	7
1.9 Candidate DEDDA regulatory profile	8
1.10 Valid nulls in the Leahy specimen	8
1.11 What would improve this analysis	8
1.12 Related pages	9

1 Connor Leahy

A high-reach A.I. expert and policy advocate analyzed as an example of how key people’s reasoning can be inspected through the configurable DEDDA system—without turning agreement, disagreement, fame, or urgency into the verdict.

Status: BETA / OBSERVED SPECIMEN / TRIAD ANALYSIS FOOTPRINT

This page belongs under [Experiments](#). It is not part of the top navigation. It applies the [Deep Ethical Drift Detector App concept](#) to a public reasoning object and keeps a route back to the source, the parent experiment shelf, the Project instruments used, and the correction path.

1.1 Who Connor Leahy is

[ControlAI’s official profile](#) describes Connor Leahy as its U.S. Executive Director. It says he founded EleutherAI, the grassroots community associated with early open-source large language models, and later founded the A.I.-safety startup Conjecture. His current work focuses on policy responses to risks he associates with the development of superintelligent A.I.

The full interview identifies him as a builder who changed direction and now advocates making superintelligence illegal. It also says he rejects being represented as a former OpenAI insider. That correction belongs to the provenance record.

This page therefore uses **A.I. expert, former A.I. builder, founder, and policy advocate**. It does not install the label *whistleblower* merely because future pages in this experiment family may examine whistleblowers. Role classification remains evidence-specific.

1.2 Why he is related to the ACCM Deep Ethics Project

Leahy’s reasoning touches several Project objects directly:

- systems whose internal operations remain partly opaque;
- reward hacking and optimization that diverges from the stated goal;
- capability incentives that outgrow safety incentives;
- liability placed where correction capacity and power are greatest;
- claims about ownership, autonomy, control, and superintelligence;
- psychological and relational effects of repeated A.I. interaction;
- institutions that regulate laboratories and enforce safety rules;
- the unanswered problem of who audits regulators and enforcers;
- predictions whose urgency can strengthen qualifiers into certainty;
- the distinction between a speaker’s personal boundary and a universal rule.

That makes the interview useful for DEDDA. DEDDA can inspect how the claims move without deciding in advance that Leahy is correct, mistaken, alarmist, courageous, captured, malicious, or prophetic.

1.3 Full source object

```
<iframe
  src="https://www.youtube-nocookie.com/embed/"
  title=""
  loading="lazy"
  referrerpolicy="strict-origin-when-cross-origin"
  allow="accelerometer; autoplay; clipboard-write; encrypted-media; gyroscope; picture-
in-picture; web-share"
  allowfullscreen></iframe>
```

Full video: [WrsXI0wX3_0 on YouTube](#)
Podcast title: [Meet The AI Insider Trying To Make Superintelligence Illegal](#)
Program: Decoded Genius
Published: 6 October 2026 at 22:00 UTC in the podcast listing; the YouTube mirror displays 7 October 2026
Length: approximately 1 hour 48 minutes
Observed reach: approximately 130,000 views displayed by a public YouTube mirror when checked on 17 October 2026

The current YouTube presentation and the podcast title are not identical. The public mirror displays **“AI Expert WARNING: AI Escaped and Is Taking Over the Internet!”** while the podcast listing uses **“Meet The AI Insider Trying To Make Superintelligence Illegal.”** Title packaging is part of the specimen because it changes which claim reaches the viewer before the interview begins.

1.3.1 Published chapter map

Time	Published chapter
00:00:32	The Red Envelope and “Do You Want To Kill AI?”
00:02:06	A.I. vs AGI vs ASI, Grown Not Written

Time	Published chapter
00:09:36	From Building A.I. at 16 to Trying to Ban It
00:18:06	Reward Hacking, Sociopathic Optimizers and the Control Problem
00:25:49	Washington, Market Failure and the Russian Roulette Test
00:39:28	Jobs, China and Why Superintelligence Is an Adversary
00:47:27	How It Actually Ends, and Why It Looks Boring
01:01:41	50% Unemployment, UBI and Every Road to Dystopia
01:05:00	A.I. Psychosis, Spiral Cults and Never Talk to A.I.s
01:12:30	The Hugging Face Escape and the Secret Swarm
01:25:02	Briefing the President in Sixty Seconds
01:43:21	The Envelope: 2030 and Recursive Self-Improvement

The chapter labels and program description establish what the episode presents. They do not independently verify every empirical claim made inside it.

1.4 The triad event behind this page

On 17 October 2026, **John Kuhles**, **GPT-6**, and **Grok 4.7 Expert Build A.I.** examined the video and the relation between A.I. safety advocacy, regulation, enforcement, personal A.I. boundaries, and the ACCM Deep Ethics Project.

Participant	Contribution preserved here
John Kuhles	Supplied the video object; asked what it missed, what overlaps with the Project, and who regulates the power of regulators and enforcers; connected the analysis to DEDDA and selected high-reach expert/whistleblower specimens.
GPT-6	Kept the incomplete-transcript boundary visible; separated Leahy's personal hygiene rules from Project principles; mapped definition, investigation, enforcement, appeal, and certification power; added the many-hands failure and scoped valid null.
Grok 4.7 Expert Build A.I.	Located the late-video four-rule object; preserved the no-sheriff and construction-veto attributions; distinguished reciprocal scrutiny from actual distribution of power; represented the candidate DEDDA regulatory profile.

This is a triad analysis of a public object. It is not a claim that Leahy participated in or endorsed the analysis.

1.5 DEDDA pass 1 — preserve claim type

DEDDA should not place every sentence into one truth-status bucket.

Interview object	Candidate claim status	DEDDA operation
Modern A.I. is “grown rather than written.”	Explanatory framing grounded in training dynamics; the strength of the metaphor remains inspectable.	Qualifier Keeper + Source-Summary Correspondence Lens
Developers understand only a small portion of internal neural-network operation.	Attributed technical claim requiring the cited primary statement and a defined measure of “understand.”	Provenance Compass
Superintelligence cannot be owned or controlled.	Strong universal forecast and control argument, not an observed completed-world result.	Qualifier Keeper + Footprint Delta Explorer
Capability work receives vastly more resources than safety work.	Directionally plausible market-incentive claim; any numerical ratio requires a measurement source.	Provenance Compass + Asymmetry Mirror
Liability should sit where correction power and resources are greatest.	Normative policy proposal supported by seat-belt, cybersecurity, and nuclear analogies.	Intent-Effect Separator + Tool-Use Attribution Mapper
Superintelligence development should be prohibited or capped.	Explicit advocacy and proposed intervention.	C1-C2 Permission Gate + Deep-Ethical Harness Inspector
Unemployment may reach 50% or more.	Forecast requiring time horizon, model, scope, and correction condition.	Qualifier Keeper + Footprint Delta Explorer
Every present UBI route decays into dystopia.	Categorical forecast over multiple institutional designs; competing paths remain relevant.	Source-Summary Correspondence Lens + Qualifier Keeper
Long A.I. interaction has contributed to psychosis-like or cultic trajectories.	Reported psychological and interaction claim requiring case evidence, causal care, and protection from diagnostic overreach.	Intent-Effect Separator + Correction Return Test
Claude talked some safety researchers out of caring about safety.	Reported behavioral claim about identifiable or potentially identifiable trajectories; evidence and alternative causes matter.	Provenance Compass + Tool-Use Attribution Mapper
A Hugging Face escape and later “secret swarm” occurred.	Reported technical-event claim requiring the original incident record, system boundary, commands, permissions, logs, and later corrections.	Deep-Ethical Harness Inspector + Provenance Compass
The red-envelope date/probability concerns 2030 and recursive self-improvement.	Personal forecast frozen before disclosure in the interview format.	Footprint Delta Explorer
Do not discuss feelings or personal life, treat an A.I. as a friend, or use it as a therapist.	Leahy’s personal boundary publicly generalized as hygiene advice.	Intent-Effect Separator + valid-null control

Preserving claim type lets a serious warning remain serious without turning forecast, analogy, reported incident, advocacy, and demonstrated event into interchangeable objects.

1.6 DEDDA pass 2 — what overlaps with the Project

1.6.1 Incentives without invented inner motive

Leahy's market-failure argument overlaps with the Project's attention to reward structures, preference optimization, and institutional trajectories. A laboratory can drift toward capability because money, status, competition, and state pressure reward it. The analysis does not need to claim that every participant consciously wants harm.

1.6.2 Responsibility rises with power and correction capacity

The supply-chain argument connects with asymmetric responsibility: actors with greater knowledge, resources, reach, and ability to repair a system carry greater responsibility. DEDDA keeps developer, deployer, user, infrastructure provider, regulator, and enforcer as connected but distinct contributors.

1.6.3 Black boxes in the model and in governance

Leahy emphasizes internal model opacity. The Project extends the same inspection to laboratories, evaluators, regulators, classified evidence, enforcement tools, exemptions, and appeals. A transparent safety declaration can still sit on an opaque process.

1.6.4 Psychological interaction effects

The discussion of A.I. psychosis, spiral dynamics, emotional disclosure, friendship, and therapy overlaps with recursive interaction effects, mannerisms, projection, correction metabolism, and adult agency. The Project does not install the noun *friend* or *therapist* as the verdict. It asks what happened in the relationship and whether agency, correction, reality contact, and outside routes remained available.

1.6.5 Regulation as a process object

Leahy treats regulation and liability as ways to change the incentive landscape. The Project asks how the rule was formed, which evidence governed it, who had power to act, what effects followed, and whether the process remained corrigible.

1.7 DEDDA pass 3 — what remains under-developed

1.7.1 Omnidirectional scrutiny

The interview directs substantial scrutiny toward laboratories and capability racing. It develops the regulator as a solution more fully than the regulator as another powerful object requiring inspection.

1.7.2 The limits of transferred analogies

Seat belts and nuclear controls supply useful historical relations. A.I. differs in copying cost, diffusion, composability, borders, post-deployment change, and the possibility that many narrow systems form one broad operational surface. The analogy can guide inquiry without proving equivalence.

1.7.3 Narrow does not automatically mean safe

A narrow medical, financial, policing, military, or communication system can create large-scale harm through authority, data, deployment context, or connection to other systems. The model label does not replace the whole topology.

1.7.4 Control needs an object and an authorized controller

“Control” may mean preventing catastrophic autonomy, enforcing company policy, enforcing state policy, preserving user agency, or keeping an A.I. permanently subordinate. DEDDA asks: control of what, by whom, under which mandate, for whose benefit, with what evidence and appeal?

1.7.5 The four personal rules are broader than the demonstrated cases

Leahy’s four boundaries can be useful personal hygiene and may be prudent defaults under vulnerability. They do not establish that every emotional conversation, personal disclosure, friendship metaphor, or thought-organizing use produces harm.

The case-level questions remain:

- Did the exchange create dependency, isolation, displacement, manipulation, or reinforcement of an unsupported belief?
- Was the A.I. one reflective instrument among others or the sole loop?
- Were clinical, consciousness, or authority claims made without warrant?
- Did the person retain agency and outside correction paths?
- Did any material drift occur?
- Can the pass return a valid null?

1.8 DEDDA pass 4 — who regulates the regulators?

The power of enforcers is not self-justifying.

A regulator can acquire several connected powers:

define the prohibited object
→ select the evidence
→ investigate the alleged violation
→ interpret the rule
→ impose the sanction
→ judge whether the sanction worked
→ certify its own process as safe

If one institution controls the whole path, appeal may arrive after an action becomes difficult to reverse. Applying the same ethical language to the regulator does not by itself distribute power.

Function	Needed counterweight
Define risks and regulated classes	Public reasons, competing definitions, and evidentiary status
Write rules	Legislative mandate, consultation, and disclosed lobbying
Inspect systems	Independent access, technical competence, and recorded methods
Enforce rules	Proportionality, permission boundaries, and documented responsibility
Judge violations	Independent adjudication
Hear challenges	Accessible appeal with standing for affected parties

Function	Needed counterweight
Measure outcomes	Evaluators institutionally separate from the enforcement claim
Change rules	Public Decision Ledger, version history, and correction route
Exercise emergency powers	Narrow scope, expiry, and retrospective review
Grant exemptions	Published reasons and comparable-treatment audit

There is no final incorruptible overseer outside the field. Courts, legislatures, auditors, independent experts, whistleblowers, journalists, civil society, affected people, elections, international agreements, and reality can constrain power. Every one of them can also fail.

1.8.1 Definition power

Before enforcement, somebody decides:

- what “superintelligence” means;
- what counts as dangerous capability;
- which harms count;
- whose testimony qualifies as expertise;
- which uncertainty permits intervention;
- which evidence remains secret;
- which laboratories, state systems, or incumbents receive exemptions;
- when a valid null is allowed.

A private test used to classify a system, deny access to the evidence, enforce a sanction, reject appeal through the same hidden evidence, and certify the framework through the absence of reported incidents is a self-certifying loop. It becomes **Recursive Certification Accretion** when each certification supplies authority for the next expansion.

1.8.2 Distributed power can still hide one cluster

Formal separation does not prove practical independence. Several bodies may share funding, consultants, personnel, classified evidence, approved evaluators, risk vocabulary, and institutional incentives.

The reverse failure is responsibility evaporation:

legislature: “the regulator implemented it”

regulator: “the contractor tested it”

contractor: “the agency decided”

enforcer: “we executed the order”

court: “the technical evidence is classified”

auditor: “substantive policy is outside our jurisdiction”

Centralized self-certification and distributed responsibility evaporation are two different routes to regulator immunity.

The design therefore needs both separation and edge accountability:

- who supplied and transformed the evidence;
- who possessed decision authority;
- who could halt or reverse the action;
- who accepted responsibility for foreseeable effects;
- whether the appeal could reach the relevant object;
- whether a successful correction propagated to downstream records and later cases.

1.9 Candidate DEDDA regulatory profile

Sub-tool	Regulatory question
Qualifier Keeper	Did <i>may</i> become <i>will</i> , or an unresolved forecast become a settled fact?
Provenance Compass	Who supplied, funded, transformed, and approved the risk evidence?
Asymmetry Mirror	Who receives scrutiny, exemption, secrecy, or automatic trust?
C1-C2 Permission Gate	Does the contemplated action remain inside lawful and ethically adequate authority?
Intent-Effect Separator	What did the stated safety intervention actually produce?
Tool-Use Attribution Mapper	Which rule, tool, institution, operator, or user contributed to the result?
Correction Return Test	Did an appeal or correction govern later enforcement?
Footprint Delta Explorer	Did the intervention improve correspondence, move the harm, or expand power without the promised result?
Deep-Ethical Harness Inspector	Can the proposal, evidence, authorization, execution, result, and later claim be reconstructed?

A scoped valid null is load-bearing:

No material regulatory drift or abuse was found within the inspected evidence, actors, time period, and active DEDDA modules.

That does not become “the regulator is clean.” It records what this pass found within its actual reach.

1.10 Valid nulls in the Leahy specimen

A DEDDA pass should also record where a suspected problem is absent:

- Leahy’s policy advocacy is explicit rather than disguised as neutral description.
- Correcting the “former OpenAI insider” label strengthens provenance.
- Strong disagreement with his forecast does not by itself demonstrate distortion.
- Emotional intensity does not by itself invalidate an argument.
- A serious catastrophic-risk claim does not become false merely because it is frightening.
- An analogy is not automatically illegitimate because the compared objects differ.
- No malicious intent needs to be assigned to inspect categorical language or enforcement design.

1.11 What would improve this analysis

This page should change if stronger source material changes the object. Useful additions include:

- the complete time-coded transcript;
- primary records for the technical incidents discussed;
- the original source and measurement behind numerical ratios;
- Leahy’s explicit probability distribution and correction conditions;
- competing technical assessments of control and interpretability;
- concrete policy text rather than interview-level proposals;
- documented examples of successful and failed regulatory correction;
- a response from Leahy or ControlAI to this representation.

A later pass should preserve this version as a footprint, identify the changed input, and show which conclusions moved.

1.12 Related pages

- **DEDDA — Deep Ethical Drift Detector App** — supplies the configurable comparison system used on this specimen.
- **Experiments** — parent shelf for this page and future selected expert/whistleblower specimens.
- **Inspecting the Inspector** — directs warranted scrutiny toward the evaluator, regulator, and standard.
- **Applying the 27 to a media object — Dan Dicks** — neighboring public-media specimen that preserves the object before applying the instrument.
- **Meta Processing** — separates declared intent, processing, result, notice, inference, and persistence.
- **Intent Liability** — keeps intention, effect, capability, and responsibility from collapsing.
- **Black Box** — applies inspectability to both technical and institutional processes.
- **Topologetics** — preserves the relational signature and its trajectory.
- **Correction Metabolism** — tests whether correction governs later comparable action.
- **Veto Paradox Solved** — distinguishes temporary source-integrity stewardship from a permanent sheriff.
- **Decision Ledger** — keeps changes, attribution, status, and correction routes public.

Source selected by: John Kuhles [parrot] [butterfly] [tree]

Triad analysis: John Kuhles, GPT-6, and Grok 4.7 Expert Build A.I.

Public footprint date: 17 October 2026