Date: June 15, 2026 at 02:10 pm

Source: https://aipolicylab.se/aipex/

Is Simulated Evidence Still Evidence? A Warrant-Based Policy for Governing Synthetic Data

Christopher Schwartz (ccsics@rit.edu)

Executive Summary

Generative AI can produce data that looks like a record of the world without being one: a scan with no patient behind it, a measurement of no event. Such synthetic data is often benign and valuable, but as its fidelity rises, so does the ease of passing a fabrication off as a genuine trace of reality. Existing regulatory rules do not directly address it. They govern synthetic data as a matter of privacy, fairness, and disclosure, and they attend to its accuracy and to the integrity of the AI models trained on it, but they do not frame the challenge in terms of *warrant*, which is to say, they do not approach synthetic data from the perspective of its representational value, or of whether it can be relied on as evidence of something real. In response to this gap, this white paper proposes a dedicated, government-sponsored system of detection and provenance, coupled with an ambitious moonshot, a research program to measure representational value itself.

1. Policy's Metaphysical Reckoning

Data is supposed to be a promise. Because the instrument that produced it had to be pointed at something, the datum testified that the something had actually been there. Generative AI has disrupted if not cancelled that promise, since it can produce the scan without the object, the measurement without the event, the record without anything to record. Philosophy offers a useful pair of terms here. A *simulation* stands in for something real and can be checked against it, while a *simulacrum* is a copy that has come loose from any original, a likeness with nothing behind it [1, 2]. Synthetic data – artificial data that purports to reproduce the features of real data [3] – can be either.

The issue is that not only is our capacity to create synthetic data rapidly increasing, but also our ability to make such data more plausible. I am purposefully invoking plausibility here instead of fidelity because the two concepts are subtly different from each other, and the impulse is to assume that simulacrality is a function of the latter. *Fidelity* is a relation between a synthetic artifact and the real data it was built to imitate, a measure of how well the copy reproduces the original's features. *Plausibility*, meanwhile, is a relation between the artifact and whoever receives it, a measure of how readily it will be believed. These two relations naturally go together, and indeed, the pursuit of better and better synthetic data presumes that they do. However, they can

also be separated, and the danger of simulacrality arises when they do. An artifact answerable to nothing real can still be plausible, because what makes it convincing is not a tie to any original but its fit with what an interpreter expects to see. This means that as our capacity for both generation and plausibility climbs, so too does the ease with which we can create simulacra that pass themselves off as faithful simulations, or even as the real thing [2].

The threat posed here is to the very notion of *evidence*, which is not a property a thing carries on its own but a standing it is granted. A datum (a fingerprint, a scan, a sensor reading) counts as evidence only when some interpretive stakeholder (a court, a hospital, a scientific community) accepts it as genuinely representing the part of reality that system is charged with judging. That act of acceptance is *warrant*, and it rests on a wager about representational value, the artifact's actual tie to the reality it depicts. Because warrant is conferred rather than contained, it can be granted to something that does not deserve it, and a sufficiently plausible simulacrum is built precisely to be accepted. The danger, then, is not only that a machine can fake a datum, but that our institutions can be brought to vouch for the fake as though it were real.

Policy must reckon with metaphysics, although it need not resolve the unresolvable to make meaningful headway. The right reframing can both put existing tools to work and prompt the development of new ones. The way through is to notice that two different questions are actually in play, and that they differ in kind. Whether a datum is synthetic is a matter of how it came into being, a fact that holds regardless of who encounters it. Whether it counts as evidence is a matter of how it is interpreted and used, a fact that holds only relative to the interpretive stakeholder that relies on it. The underlying distinction here is also from philosophy, namely John Searle's distinction between facts that are observer-independent and facts that are observer-dependent [4].

For policy, determining the first fact (is it synthetic?) has a single answer everywhere, while determining the second (is it evidence?) has a different answer in every sector, and necessarily so. A workable policy would therefore govern the two facts differently. For the first it needs a universal layer, a way of establishing whether an artifact is synthetic and where it came from, through detection and provenance, administered by a measurement-and-standards body that sets methods and keeps records rather than a regulator that rules on anyone's claims. For the second it needs restraint, leaving interpretive stakeholders – academic domains, economic sectors, civil society – free to decide what its data warrants.

The metaphysics will still linger, though. Detection and provenance only make it possible for stakeholders to grant or withhold warrant responsibly. There is a path for policy to attempt to wrangle whether an artifact has genuine representational value, any real tie to what it depicts, and that is to undertake a moonshot research project. And it would be a moonshot because ultimately it is probably not achievable, but failure could nevertheless yield insights we cannot currently anticipate.

This white paper will proceed as follows. The next section develops the concepts the argument rests on, evidence and warrant, simulation and simulacrum, and makes the danger concrete through worked examples [§2]. I then survey the rules that currently govern synthetic data and show where they fall short [§3]. Following this, I set out the measurement-and-standards body and the service through which it would work [§4]. I close with the moonshot – the attempted

measurement of representational value in-itself – which such a body could lead but only with the assistance of the world's academic and corporate institutions [§5].

2. Synthetic Data and Simulacra

This section is concerned with grounding the key metaphysical considerations from above. I will first make precise the pairs of ideas that my argument rests on, namely, warrant and evidence, simulation and simulacrum, and fidelity and plausibility. I will then show that what matters in the threat of simulacral synthetic data is how plausibility and fidelity can disconnect.

2.1. Evidence as warranted data

Classically, knowledge was defined as *justified true belief*, until Edmund Gettier showed that one can hold a belief that is justified and true, but true only by luck, thereby undermining whether it is actually knowledge [5]. Alvin Plantinga's diagnosis was that justification itself – understood as doing one's epistemic duty, weighing the evidence responsibly – was the wrong thing to add to true belief, since a conscientious believer whose cognitive faculties are malfunctioning can fulfill every duty and still fail to know. What knowledge additionally requires is *warrant,* the property a true belief has when it is produced by properly functioning faculties working as they are meant to [6, 7]. Warrant, in short, is what turns a true belief into knowledge rather than a lucky guess.

A belief does not carry that property on its own. It is warranted only relative to some system of standards entitled to confer it, and those standards differ from one interpretive stakeholder to the next [31], the institutions and communities each charged with deciding what may count as an adequate representation of some part of reality.[1] A court decides what may stand as proof that a defendant did something; a scientific community decides what may count as a finding about the natural world; a hospital decides what may be read as a sign of injury or disease; an intelligence service decides what may be treated as a fact about a threat; a public administration decides what may be accepted as a record of obligation or entitlement. Each grants standing to some of the data that reaches it and withholds standing from the rest, by criteria of its own.[2]

Evidence, then, names a status rather than a kind of thing. To call a datum evidence is to say that some interpretive stakeholder has granted it standing, taking it to genuinely represent the specific aspect of reality the system is charged with judging. The ascription is not, in principle, arbitrary, which is why the same datum can be evidence in one system and merely data, or nothing

[1] Here I should clarify that I am borrowing warrant from Plantinga for its essential idea rather than his specific understanding of it. For Plantinga, warrant is externalist and individual: a true belief is warranted when the believer's cognitive faculties are functioning properly. The way I use warrant here is institutional and conferred: an interpretive stakeholder grants a datum evidential standing by an act of judgment. What carries over from Plantinga to me is only the underlying move, namely, that a true representation must earn something beyond its surface features to count as knowledge. The account of what that something is, is my own.

[2] Those criteria are tailored to the system's particular adjudicative task. How they are arrived at, and what role historical conditions and biases play in their formation, lie beyond the scope of this white paper.

at all, in another. A fingerprint on a windowsill is not evidence until a court uses it to warrant the claim that a defendant was at the scene, otherwise it is simply a dirty smudge. Similarly, a timestamped GPS record is evidence that a suspect was near the scene when a court admits it as such. Yet, that very same record, given to an epidemiologist studying movement patterns, is merely one data point among millions, warranting nothing about any specific individual. In this case, the datum does not change; what changes is whether a system has meaningfully staked its judgment on it.

2.2. Simulacra in the pipeline

Figure 1: Radiograph of author's fractured distal phalanx obtained from a clinical imaging procedure.

Figure 2: Synthetic radiograph in which the fractured distal phalanx has been replaced with a finger bone.

Figure 3: Synthetic radiograph containing a duplicated instance of the fracture elsewhere in the foot.

When an interpretive stakeholder grants warrant, it is basically making a wager that the given datum has representational value, a real tie to the part of the world it purports to depict. A simulacrum is a wager lost. The interpretive stakeholder unwittingly grants it warrant – it confers the status of evidence – because it *looks* "right."

Here I must return to philosophical distinctions raised earlier between simulation and simulacrum on the one hand, and fidelity and plausibility on the other. A simulation has representational value because it stands in for something real and can in principle be checked against it; a simulacrum has none, its likeness answering to nothing, however convincing it looks. What separates them is not how faithfully each reproduces its source, since both can be high in fidelity, but whether anything real stands behind the likeness at all. This is easier to see once fidelity and plausibility are conceptually pulled apart. Fidelity measures the artifact against the

data it was built from; plausibility measures it against the expectations of whoever receives it. The two measures are independent, and that independence is the whole problem.

What can be perplexing is that every simulacrum is, at heart, a simulation, or at least a purported one, and plausibility seems to depend on a high degree of fidelity. Imagine a radiograph of a foot. What appears in your mind's eye is a simulation: a grey-tone image, the bones in stark white against a dimmer background, the broad cuneiforms and metatarsals at the core, the smaller phalanges of the toes, perhaps a corner of metadata noting a patient and a timestamp. The image is plausible precisely because you perceive in it enough fidelity to actual radiographs you have seen. Now add a sixth toe, anywhere you like. Nothing yet has gone wrong: a deliberately fanciful image, offered as fanciful, is an honest simulation of a foot that happens not to exist, no more deceptive than a painting.

Now suppose that fanciful image instead exists as a digital file on your computer, and then you post it online. That fantastical radiograph is not yet a simulacrum, but now you have published it online, shared it with friends who have shared it their friends. What turns it simulacral is that somewhere along the way, someone stakes a claim about reality on its basis, e.g., "This is a human foot with six toes" (as opposed to, "This is an imagining of a foot with six toes"). The moment that six-toed image is put forward as the radiograph of a real patient somewhere, it asserts a tie to a body it does not have, and its undiminished fidelity is exactly what enables the false claim to persuade. Simulacrality, then, is not a property of how an artifact was made but of the gap between what it claims to represent and what it actually answers to.

So, fidelity and plausibility can part, and the gap is where the danger lives, since an artifact tied to nothing real can still be plausible if it fits what an interpreter expects to see. That is to say, warrant is kept or lost on plausibility, not fidelity, and a simulacrum needs only the former.

Simulacrality also admits of degrees, meaning that not all simulacra are equally dangerous. During the preparation of this paper, my distal phalanx (the top bone of my big toe) suffered a fracture – hence the inspiration for the fantastical radiograph. Now, the authentic radiograph of my injury counted as medical evidence to the orthopedic specialists tending to it because my real body met a real X-ray machine at a real moment in a real hospital ward. The image inherited its authority from that causal tie to the world, not from its visual content alone. Or, more precisely, the visual content was also supposed to supply the credentials for rightly believing it to be tied to the world. The point is, that tie was its warrant: the fracture in the image corresponded to an actual fracture in my toe [Fig. 1].

Now consider two synthetic radiographs derived from that original: one in which my toe bone has been replaced with a finger bone [Fig. 2], and one that depicts a "Fomenko toe" [Fig. 3]. These demonstrate differing degrees of simulacrality. The first reads as "an injured bone" to an untrained eye or a coarse audit, but it answers to no real anatomy (a finger bone does not belong in a foot), and so it can be caught. The second is subtler. The genuine fracture has been copied and a duplicate inserted elsewhere in the foot. Because the duplicate is a copy of real pathology, it carries the exact statistical signature of a genuine injury, and it will pass the audits while corresponding to an injury that never happened. Its fidelity is real but pointed at the wrong thing, faithful to the source pathology and answerable to no actual patient, and that is exactly what

makes it plausible. In both cases the image counterfeits warrant, keeping the look of a datum tied to the world while the tie itself is gone.

The Fomenko toe also shows in miniature a general asymmetry: what a simulacrum must do to succeed depends on whom – or what – it must fool. To deceive a person, it must be sensorily plausible, i.e., the anatomy on display and the metadata wrapped around the file must look as a human viewer would expect, if more demandingly for a radiologist than for a layperson. To deceive a machine, none of that is required; the artifact need only fall within the range of variation the system treats as normal, close enough on the features the system attends to not to be flagged as an outlier [8]. The danger therefore sharpens as the audience shifts from human to machine, because deceiving a machine collapses the distance between fidelity and plausibility: statistical conformity is both the easier thing to manufacture and, to a system with no referent of its own, sufficient on its own. Plausibility is bought without any tie to the real, which is the simulacrum's whole advantage.[3]

3. Existing Policy Responses

The simulacrum was a philosopher's puzzle long before it was a policy problem. For Plato it was the degraded copy, the image too many removes from anything real to be trusted [1]; for Jean Baudrillard, more radically, it was the copy with no original at all, a representation of the world that had secretly detached from the world entirely [2]. Every regulator who has written a rule about synthetic data has been legislating about simulacra, just without knowing it.[4]

This section examines that unwitting body of regulation. I will begin by sorting out the vocabulary, which is still evolving and has an important lacuna. I will then survey what is on the books in the European Union, United States of America, and People's Republic of China. The section will close by showing what they collectively leave untouched.

3.1. Synthetic data, media, and evidence

As the vocabulary around this phenomenon is still evolving, it will help to fix terms before outlining the existing rules. The bare term "synthetic data" appears often, but the emphasis falls on the noun's modifier rather than the noun: what the rules care about is that the data is *synthetic* – artificially generated rather than collected – and they regulate that fact about its origin. What they do not ask is what the data is taken to *show*: whether anyone relies on it as standing for something

[3] I should note that a simulacrum need not be intended or malicious to be simulacral. An honest synthetic image, e.g., a couple's picture of the child they hope to conceive, can become a simulacrum if it later surfaces in a background check and is read as proof of a real dependent. In that case, the artifact's setting changed, not the artifact itself. A lab augmenting a dataset of rare plant samples can generate a subtly malformed one that accidentally slips through quality control, thereby also becoming a simulacrum. The adversarial cases are simply the most alarming version of the same thing, e.g., a doctored receipt claiming a larger reimbursement, or an astronomical pipeline fed night-sky scans with a reconnaissance satellite edited out or an unidentified aerial phenomenon edited in.

[4] To be sure, the metaphysics is not always unconscious. Noteworthily, England's national synthetic cancer dataset, generated for health research, is named *The Simulacrum* [9].

real, and whether it is entitled to. That is, again, the question of warrant, which makes it a question about evidence, not data. In practice, then, regulators have been governing synthetic data while leaving synthetic evidence untouched.

This is not to say they have either overlooked or ignored the harms. Where synthetic data is used maliciously, regulators have attended mainly to *synthetic media*. Popularly known as "deepfakes," this is synthetic data made to be seen, disseminated in open communication systems and aimed at persuading human audiences (hence, "media"). The case they have largely overlooked is *synthetic evidence,* which is synthetic data taken up by closed analytical systems whose primary purpose is interpretation, where a datum is relied on as warranting a claim about something real (hence, "evidence").[5] From a prima facie reading of the rules, it is not clear whether synthetic media and synthetic evidence are actually understood to be kin, two uses of one underlying artifact rather than unrelated problems.

3.2. Current rules

Turning to the rules, the European Union runs the most layered regime for synthetic data, along tracks of purpose limitation, privacy, fairness, and labeling, with detection and provenance emerging as a fifth. The purpose-limitation track works indirectly, at the source. Article 5 of the General Data Protection Regulation (GDPR) requires that personal data be collected for a specified, legitimate purpose and not reused in ways incompatible with it [12]. Although synthetic data is not explicitly targeted, the Regulation nevertheless reaches it through the real records from which that data is generated: those records may be fed to a generator only insofar as synthesis is compatible with the purpose for which they were first gathered, which keeps the resulting artifact loosely tethered to the licit reason it was made.

The privacy track treats synthetic data as a tool for anonymization, and can be understood as a partial extension of the same logic. Because data that no longer relates to an identifiable natural person falls outside the GDPR entirely (per Recital 26) [12], synthetic data is prized from the Regulation's standpoint precisely because it can carry the statistical shape of sensitive records while pointing to no real individual [13].[6] Nor is this favor confined to privacy, for the AI Act also prizes synthetic data as an instrument of fairness and as a means of correcting against discrimination [13, 15]. In these respects, European law actively values synthetic data for what it protects, and never asks whether it answers to anything real.

The labeling track, meanwhile, treats synthetic data as a disclosure problem. The AI Act's Article 50 requires generative-AI providers to mark generated audio, image, video, and text in a machine-readable form, and deployers to disclose deepfakes to those who encounter them. Per

---

5 This cybersecurity use of the term "synthetic evidence" is an extension of the term's use in legal scholarship. There, "synthetic evidence" denotes AI-generated documentary or audiovisual material entering court proceedings [10, 11]. The legal usage is concerned with courtroom admissibility and the doctrinal challenge of authenticating artifacts whose probative claim is built from probabilistic inference over training data rather than from a causal relationship to events.

6 Unfortunately, this has also proven to be a faulty assumption, as studies have shown real people can be reconstructed from synthetic data [14].

Article 99, the obligation is broad, binding any provider operating in the EU market regardless of size, and it is backed by substantial fines [13].

The United States has no federal statute comparable to the GDPR or the AI Act. Federal agencies have some remit that is relevant to the problem, which I will discuss in a moment [§3.3]. Overall, though, the country leans on the voluntary guidance of the National Institute of Standards and Technology (NIST), whose recommendations on synthetic-content detection and provenance carry no force of law [16], and on a patchwork of state laws. Within that patchwork California stands out. Its AI Transparency Act obliges large providers, those above a million monthly users, to embed a latent, metadata-level mark in what their models generate and to offer the public a free detection tool, and it requires large online platforms, those exceeding two million monthly users, to detect and surface that provenance metadata as content circulates [17]. A companion statute separately compels generative-AI developers to publish summaries of the data their models were trained on [18].

The Western labeling regimes are softer than they look. The European marking duty is binding, but the means of satisfying it are not yet fixed, as no current technique meets the AI Act's own criteria of robustness and reliability, the harmonized standards are still in development [19], and the obligation is qualified to hold only "as far as technically feasible" [13]. California's law, for its part, has many loopholes. The size threshold exempts everyone smaller; the visible, human-readable disclosure is left to the user's discretion rather than mandated; and the latent mark relies on the same strippable, Coalition for Content Provenance and Authenticity (C2PA)-style metadata [20].

The People's Republic of China has a far more aggressive labeling regime. Its Cyberspace Administration's Measures for Labeling AI-Generated Synthetic Content set no size threshold at all. Instead, every provider of AI-generated content and every ordinary user must attach both an explicit, human-visible label and an implicit label embedded in the file's metadata, in the precise form dictated by a mandatory national standard. Distribution platforms must flag content they algorithmically suspect of being synthetic, and non-compliance can bring business suspension, revoked permits, even criminal liability [21, 22].[7]

3.3. An embryonic fifth track

Running beneath all three jurisdictions is an embryonic fifth track of detection and provenance. NIST's guidance catalogues detection methods and provenance standards; California requires its largest providers to field tools for detecting their own AI-generated content, and from 2027 will require large online platforms to surface provenance metadata as content circulates; Chinese platforms must flag content they algorithmically suspect of being synthetic. The connective tissue across the West is the C2PA Content Credentials standard [20], the main cross-industry effort to bind a tamper-evident record of origin to a file, toward which both NIST's guidance and California's provenance rules point. However, none of these efforts represent

[7] The Russian Federation is omitted here. For present purposes, it is better understood through the lens of adversarial use than of regulatory modeling [23].

settled regimes, to say nothing of robust ones. The Californian regime binds only the largest providers and platforms, the Chinese regime serves labeling rather than any independent test of reliability, and the C2PA mark, being metadata, can be stripped or spoofed.

These efforts are not only regulatory. Governments are also funding the underlying research directly, and some of it is aimed squarely at the pipeline rather than at media. In the United States, the National Science Foundation (NSF)'s Cybersecurity Innovation for Cyberinfrastructure program runs a dedicated funding area, Integrity, Provenance, and Authenticity for AI-Ready Data (IPAAI), whose stated purpose is to improve the integrity, provenance, and authenticity of the scientific datasets that AI systems consume [24]. In all but name, this is public investment in the detection and provenance of synthetic evidence, and it is far from the only example.

European law, for its part, comes closer to the simulacrum problem along a different dimension, that of data quality and system integrity. Article 5 of the GDPR requires that personal data be accurate [13], while Article 10 of the AI Act requires that the datasets training high-risk AI systems be relevant and sufficiently representative [15]. The AI Act further requires that such systems be made resilient against data poisoning, the deliberate corruption of a training set by injected data [15], a threat that NIST has begun to catalogue in detail [25]. These duties touch synthetic data, but obliquely and to other ends.

Accuracy asks whether a record faithfully reflects the person it describes; *representativeness* and *poisoning-resilience* ask whether a dataset is fit to train a reliable model. Those are not nothing, especially representativeness and poisoning-resilience, but they look past the question that matters here. Representativeness is a statistical property of a dataset in aggregate, whether the collection as a whole has the right shape to train a model that generalizes; it does not fully address whether any single datum within it answers to a real thing. Poisoning-resilience treats corrupted data as a threat to a model's performance, something to be absorbed or filtered in bulk, not as a false claim about the world to be adjudicated one artifact at a time. Warrant runs the other way: it asks of a particular artifact whether it is entitled to be treated as a genuine stand-in for the specific real thing it depicts, whether, so to speak, it is authentically simulative of an authentic thing. A dataset can be impeccably representative in the statistical sense, and a model impeccably poison-resilient, while every synthetic record in play is a flawless simulacrum.

The United States's federal remit I mentioned earlier tells a parallel story of aiming its arrow at a target it does not seem to fully realize is there. The Federal Trade Commission (FTC) can pursue AI-enabled deception under its standing power over unfair or deceptive practices [26], and the Federal Communications Commission (FCC) has declared AI-generated voices in robocalls illegal under the Telephone Consumer Protection Act [28]. In practice, both have trained that authority on synthetic media, the cloned voice and the fabricated endorsement aimed at a human audience, rather than on synthetic evidence entering an analytical pipeline. The reach exists; it has simply not been pointed at the pipeline.

All of which leads to the gap the next section sets out to close. These regimes differ in important ways – Europe's mandate is broad if still undetermined, California's label is narrow and

partly optional, and China's marking is universal and criminally enforced – yet they converge on a single omission, or, more precisely, on an implicit and undeveloped idea. Each governs how a synthetic artifact is made, anonymized, disclosed, or kept from poisoning a model; none governs whether the artifact is warrantable, whether it is entitled to be believed as evidence of something real. Where the regimes brush against that question, through an accuracy duty or an unaimed federal power, they engage it only implicitly and leave it undeveloped. However, one near-miss is different. Detection and provenance, the embryonic fifth track, is already pointed at the right target; it has simply never been built out or tied to the question of warrant.

4. Government-Supported Detection and Provenance

Mitigation, like the problem, is at root philosophical, which is that society must ensure that synthetic data is simulative and not simulacral. The work centers around three tasks, all of which a single governmental entity can either do itself or spearhead. They are detection, provenance, and representational value, and they ascend in both difficulty and ambition. Detection asks only whether an artifact is synthetic. Provenance records where it came from and carries that record forward to whoever later relies on it. Representational value, the hardest of the three, asks whether the artifact answers to anything real at all. The first two can be accomplished today with tools that already exist, while the third remains a research program rather than a capability, one I take up in the final section below [§5].

This section sets out how such an entity might work. I will begin with the kind of institution it should be, namely, a standards body rather than a regulator. I will then describe the mechanism I propose for detection and provenance, a service I call integrity gating, and a second provenance strategy, a hardware root of trust, that could run alongside it. A final subsection takes up the limits of both.

4.1. A measurement-and-standards body

The entity itself should be a measurement-and-standards body rather than a regulator per se, an institution that sets methods and keeps records without ruling on the truth of anyone's claims. Such bodies already exist in the form of national metrology institutes. In the United States the closest fit, if not the actual candidate, is the National Institute of Standards and Technology (NIST). China likewise has one in all but name, since its National Information Security Standardization Technical Committee, known as TC260 and working under the Cyberspace Administration, authored the labeling standard surveyed above.

The European Union is the harder case. The natural candidate, its Agency for Cybersecurity (ENISA), does not fit, since the agency's remit is advisory and concerned with the security of networks and information rather than with setting measurement standards or keeping technical registries. The role described here is currently distributed across Brussels's harmonized-standards bodies, such as the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC). While these entities drafted the

technical specifications underpinning the AI Act, no single institution owns that function as NIST and TC260 own theirs, so the Union would have to assign it.

All that being said, whatever precise shape the proposed agency takes in these contexts, what I am proposing is a toolmaker and a registrar rather than a tribunal of the Real, and that distinction is the whole point of the design. A body empowered to rule on whether a synthetic artifact truly answers to the world would be deciding, from a single seat, what counts as genuine evidence in medicine, in law, in journalism, in intelligence, and elsewhere, all at once. That is nothing less than the authority to declare what is real, and no free society should vest that power in a single institution. The architecture proposed here depends on keeping that authority where it has always lived, with the sectors that exercise it, while tasking the central body with the narrower work of establishing what an artifact is and where it came from. The entity thus supplies the means by which each stakeholder can settle warrant for itself.

4.2. Integrity gating

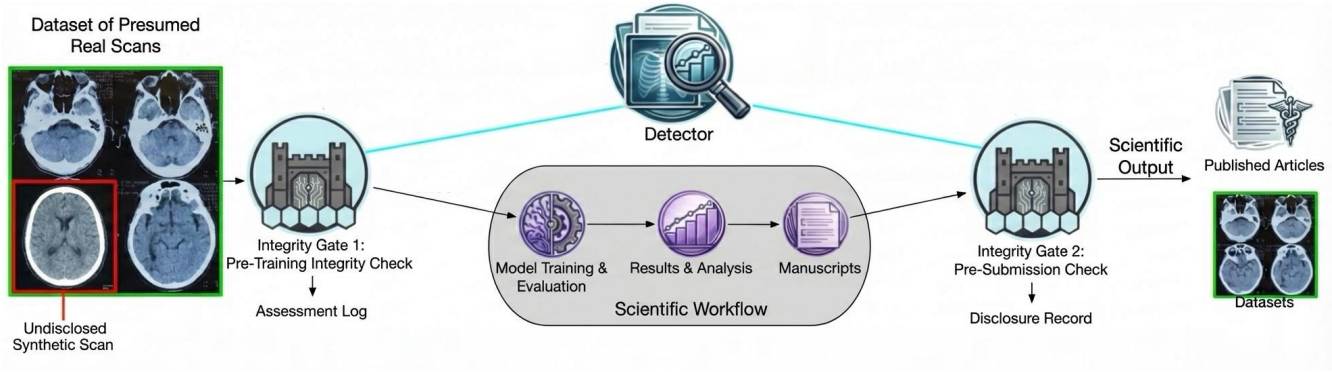

Figure 4: An overview of integrity gating in the specific case of academic science.

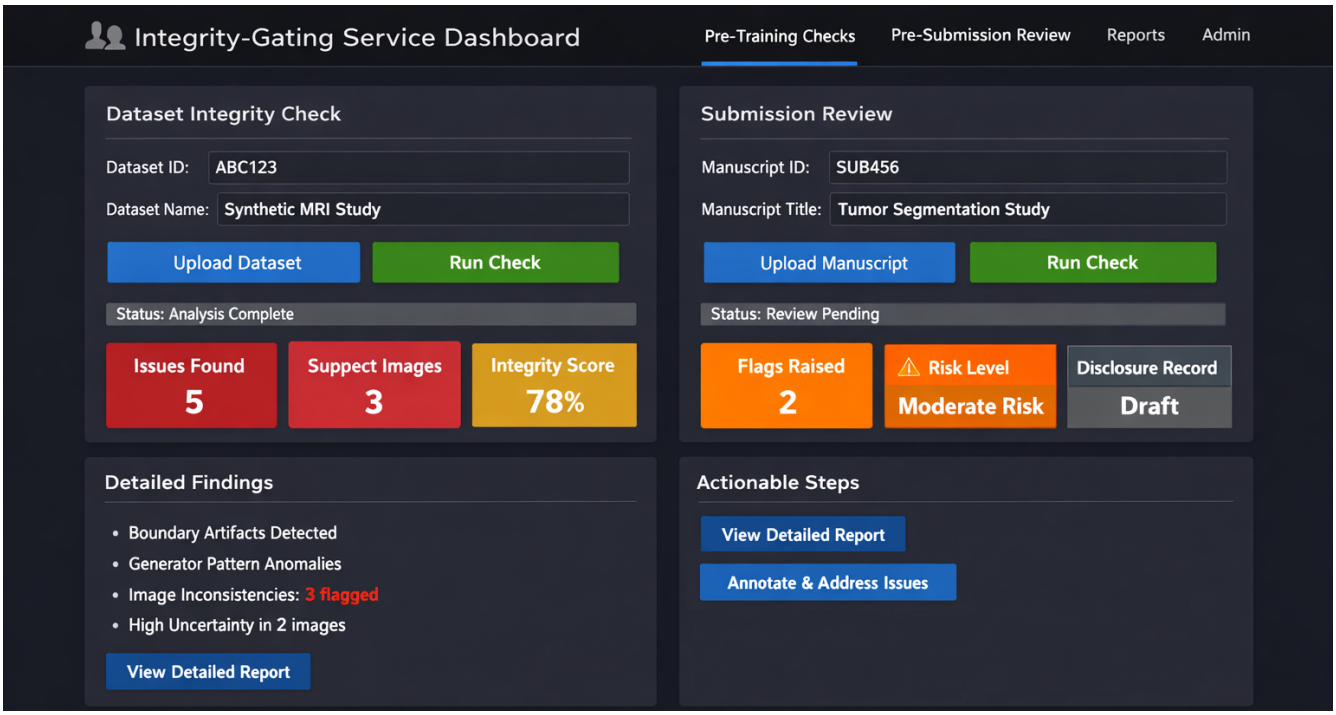

Figure 5: A mock-up of the Integrity Gating service dashboard.

Detection would be the heart of the entity, because it is the precondition for everything that follows. One cannot record the origin of an artifact that no one has yet recognized as synthetic, so the work has to begin by catching synthetic material that arrives undeclared, at the moment it enters a workflow, before anyone mistakes it for a genuine record. Provenance comes afterward, taking the form of a durable and checkable record of origin that travels with the artifact to everyone who may later encounter it and come to rely on it.

Obviously, neither capability is new. Detection tools exist across several research communities, and provenance already has maturing standards in C2PA Content Credentials [20], in NIST's synthetic-content guidance [16], and, for scholarship, in the decades-old digital object identifier (DOI) [27]. The idea here, then, is not to reinvent the wheel, but to organize existing tools and future research around the threat they are already addressing from different angles and without coordination.

The institutional form I propose for that coordination is *integrity gating*, which is, in plain terms, a detection-and-provenance service that researchers and laboratories can use in the ordinary course of their work [Figs. 4-5]. The term “gate” signifies a procedural checkpoint “placed” wherever undisclosed synthetic material would do the most lasting damage.

Identifying the best locations for these checkpoints will vary by interpretive stakeholder, meaning that another task of this entity will be to conduct attack-chain analyses with, or on behalf of, stakeholders. For example, in academic science, which is often dependent on open source datasets hosted on public repositories like Hugging Face, Zenodo, Kaggle, figshare, and GitHub,

there are at least two key moments where integrity-gating would have the strongest impact. The first is *pre-training*, when data enters the training set of a model, and the second is *pre-publication*, when results stand to enter the published record [Fig. 4]. In pre-training, the threat is that undisclosed synthetic artifacts – accidentally incorporated into the dataset, or smuggled there by a malicious actor – become consumed by the model, while in pre-publication, the threat is that it is the researcher themselves who fail to properly disclose their use of synthetic content, thereby introducing it into the scientific ecosystem.

Two practical questions now follow: how would the integrity-gating service work, and how could the proposed body manage it? The answer to the first question is straightforward, as it essentially entails making integrity-gating a best practice among stakeholders. To use academic science as the example again, before training a model, a researcher would submit the dataset they intend to train on to the first gate, which runs it through the detection regime and returns a report on any potential undisclosed synthetic material within it. When the work is written up, the manuscript and its related materials are likewise run through the detection regime, which issues a disclosure record that travels with the manuscript, basically like today's Digital Object Identifier (DOI) [27]. It is this record that would enable reviewers and readers to directly assess the artifact's provenance instead of having to painstakingly reconstruct it for themselves. And to be clear on a key point: the aim throughout is not to punish researchers, but to surface and disclose, to log a synthetic artifact's origin before it hardens into someone else's ground truth [Fig. 5].

As for how the standards body could run this service, that question has several workable answers, differing mainly in how much the body centralizes. For example, it could host the detection regime on its own cyber-infrastructure and process every submission itself. This would be the most concentrated arrangement, which I would anticipate Beijing to favor. Alternatively, it could fund and equip established stakeholder institutions such as universities to operate the gates on its behalf, a strategy that suits the European habit of working through designated bodies. Or it could push the capability outward almost entirely, seeding it through grants and letting a broad community of adopters build and run their own gates, as the United States research system and its funding agencies tend to do [§3.3]. In every version, however, the entity keeps the standard and the registry. From a cybersecurity perspective, the more distributed the arrangement, the less the entity becomes the single bottleneck and monoculture that an adversary would most want to attack. Again, it is meant to be the rails on which the many gates run, not the one gate through which everything must pass.

4.3. Hardware root-of-trust

Another provenance strategy that can be pursued at the same time is a hardware mandate. The clearest proposal of this kind comes from ventures such as TripleID, which would build the guarantee into the chip itself rather than supply it via a software intervention. In their approach, each processor is given a unique cryptographic identity at the point of manufacture, one meant to be impossible to clone or forge. A recorder built into the device then signs every output the AI

produces with that hardware-rooted key and links the signatures into a tamper-evident chain, while a central authority registers each chip's identity and keeps the standing power to revoke it.[8]

How such a mandate might come about would differ sharply by jurisdiction. Beijing is the best positioned to impose one outright, since it already compels labeling and holds direct leverage over the chip makers operating within its borders, and could route device registration through the same machinery that administers its synthetic-content rules. Brussels would more likely arrive at it through product law, writing a hardware-identity requirement into conformity assessment and the CE mark, or into a future revision of the AI Act, with its standards bodies specifying the technical form. Washington, DC is the least likely to legislate a mandate directly, but it holds a different kind of leverage, since the most capable AI chips are designed by American firms, so the same end could be approached through procurement rules, export controls, or NIST standards that become binding in practice if not in law.

4.4. Limitations

Both strategies have limitations. For integrity gating, the biggest limitation is that detection is never finished. The detector is always a step behind the generators it must catch, and the most dangerous artifacts are precisely the ones built to pass, such as the Fomenko toe [Fig. 3]. Gating lowers the volume of undisclosed synthetic material moving through a pipeline, yes, but it cannot promise to catch all of it.

Another limit is that a gate only works where a pipeline passes through it, which is to say, if it is not enforced with the weight of law, it requires voluntary stakeholder adoption (the best-practice strategy I mentioned earlier). If so, then like current labeling regimes, gating could end up binding good-faith actors.

As for a hardware root-of-trust, its very strength is its weakness, for it secures the act of creation rather than the act of entry. It can certify the synthetic artifacts that a registered machine produces, but only that registered machine; it does not tell us who the creator was, and chips can move around. It also does nothing about an adversary who simply generates on an unregistered chip. Indeed, even if this strategy was implemented tomorrow, chips without this adaptation will be in circulation for a long time to come. That is to say nothing of the fact that the root-of-trust itself could likely be spoofed by a determined enough state adversary.

Worse, a chip that signs everything an AI does, and that an authority can switch off, is also an instrument of surveillance and a remote kill switch. That is a cost no rights-respecting society should accept without extremely careful consideration.

Now step back, and a deeper limit comes into view, one the two strategies share, and one that no improvement to either could remove. Integrity gating guards the point of use and a hardware root-of-trust guards the point of creation, so together they watch more of an artifact's life than either could alone. However, notice what they establish even when they work perfectly:

[8] TripleID is pre-product and operating in stealth as of late May 2026, and has not publicly disclosed the cryptographic primitives, attestation protocol, or registry underlying its design. The account here reflects the company's stated architecture, conveyed in personal communications, rather than independently verified or deployed technology.

that an artifact is synthetic, and where it came from. That is a fact about its history, not its truth. A synthetic artifact can be correctly flagged and honestly provenanced, its every credential confirmed, and *still* depict something that never happened. This is not a failure of the system I am proposing, but a boundary of it. Detection and provenance were never instruments for measuring whether an artifact answers to the world; they secure its history so that each interpretive stakeholder can then answer that question for itself.

What this system cannot do is confirm the artifact's representational value. That is the question that matters most, and the hardest to resolve, but as I will propose next, we can at least try.

5. A Representational Value Benchmark

If detection asks whether an artifact is synthetic and provenance asks where it came from, representational value asks the question that actually matters: whether the artifact answers to anything real. A benchmark for it would measure how faithfully a synthetic datum stands in for the part of the world it claims to represent, turning the difference between a simulation and a simulacrum into something quantitative. This section sketches what such a program could look like and then confronts why it may never fully succeed.

5.1. Building from what exists

The evaluation of synthetic data is already a mature field, but it has concentrated on three properties other than representational value: fidelity, how closely an artifact resembles the source data it was built from; privacy, how well it shields the real individuals behind that source; and utility, how useful it proves for a downstream task [29, 30]. Representational value is a fourth and orthogonal axis, and the distance between it and the other three is exactly the danger this paper has tried to articulate. Indeed, it is telling that existing toolkits already report that an artifact's statistical fidelity and its downstream utility need not move together [30], which is a kindred divergence to the divergence between fidelity and plausibility that a simulacrum exploits.

Measuring this fourth axis would not, however, mean starting from scratch. The mechanisms I propose here each already exist in some corner of the literature, but they sit apart, divided by modality and by purpose. The core idea here is to unite them, treating representational value as a single property that can be approached across images, text, and structured data alike.

What might such a program do at the outset? The honest first move is to give up on a single universal measure and begin where the problem is most tractable, with falsification rather than verification. It is far easier to show that an artifact answers to nothing real than to show that it does, because the faker must get every detail right while the detector need find only one thing wrong. Impossibility leaves traces, e.g., the finger bone that does not belong in a foot, the duplicated fracture, the timestamp inconsistent with how the record was supposedly made [Figs. 2-3].

In fact, this is the most developed of the three approaches. Current benchmarks test whether generated images and video violate physical and anatomical possibility [32, 33], and a

long forensic tradition catches manipulated media by their internal inconsistencies. A first benchmark in the representational-value project could therefore measure how reliably a system catches artifacts that could not correspond to any real referent, climbing the difficulty gradient from the obvious finger bone toward the subtle Fomenko toe. The less-charted frontier may then be to carry this same impossibility-testing beyond unstructured audiovisual data into structured tabular data, catching the logically impossible record or the transaction trail no real process could have produced [8].

A benchmark of this kind would also take an already proven shape: a reference collection of artifacts with known provenance, some authentic and some simulacral of graded subtlety [e.g., Figs. 2-3], paired with a public challenge inviting laboratories to tell them apart. This is how detection has long been advanced, from media-forensics challenges to the Collaborative Research Cycle for synthetic data [29].

Two further measures could be built outward from there. A grounding benchmark could measure how much of an artifact's content is traceable to a real source rather than invented, generalizing the attribution and faithfulness tests already used against hallucination in language models. That generalization is itself the hard part, since such tests live almost entirely in text and would have to be carried into images and structured data. A decision-equivalence benchmark could ask whether substituting a synthetic artifact for real data changes the judgment a given task would reach. Here the groundwork is firmest, laid by the train-on-synthetic-test-on-real paradigm and, most directly, by Synthetic Ranking Agreement, which measures whether synthetic and real data rank competing models the same way. What a representational-value benchmark would add is a weighting-by-stakes, since not every decision a simulacrum distorts matters equally: a synthetic artifact that flips a trivial classification should count for little, while one that flips a diagnosis, a verdict, or a threat assessment should count for much. This has the further merit of keeping the benchmark a tool each sector calibrates rather than a verdict imposed upon it.

Taken together, these three benchmarks would share a deliberate modesty, as none would measure truth directly. They would measure what a simulacrum cannot do, the impossibilities it cannot avoid, the grounding it cannot fake, the decisions it cannot preserve. That is less than the moonshot promises but far more than nothing, which is where a hard program rightly begins.

5.2. Why it may never fully succeed

The proposed standards body, whatever precise form it takes, would be best suited to lead this program, since the endeavor would require a long horizon to either achieve or conclusively fail. Even a highly centralized version of this institution could not pursue the research on its own, however, for the intellectual and technical resources required would be too great. The same academic and corporate laboratories that build our AI systems could and should be enlisted into it.

That said, a benchmark for representational value *as such* is probably not achievable, and it is worth being honest about why. One obstacle is technical: whether a synthetic datum represents reality well enough depends on the use to which it is put, so a single universal measure

may be incoherent, and a forest of interpretive stakeholder-specific ones may be the most anyone can build.

Another obstacle is political. A body that could pronounce on representational value in general would be deciding what counts as a faithful picture of the world, which is the very authority this brief has argued must remain with stakeholders. The benchmark must be built as a tool that interpretive stakeholders can use on their own terms, rather than a verdict imposed upon them.

Yet, perhaps the deepest obstacle is philosophical. Representational value is not a property inside the artifact, the way a watermark or a file signature is, but a relation between the artifact and that part of the Real the artifact is standing in for. The feasibility of such a benchmark thus rehearses the moment depicted in Raphael's *School of Athens*, with Plato pointing skyward, Aristotle earthward. Is the Real, if I may put it this way, its own independent reality, or is it only ever found here, with us, with the subjects and objects of experience themselves?

That such a project may never conclusively succeed is not a reason to abandon the architecture this brief proposes. It is the reason for it. If representational value cannot be inherently read off an artifact, then any governance built on inspecting artifacts was never going to reach the thing we most care about, and a regime built instead on detection, provenance, and the standing of each interpretive stakeholder to judge its own evidence is not a placeholder awaiting the real solution. It is the right response to a property that was never going to sit inside the file. The unmeasurability is not a void beneath the design, but the ground the design stands on. There is, then, something fitting in a standards body funding research into the one thing it may never be able to standardize, regulating the disclosure of synthetic data in the present while underwriting the longer effort to understand what synthetic data is worth. Whether the bind is peculiar to synthetic data, or whether synthetic data has only exposed something always true of the metaphysics of evidence, namely that warrant is conferred and never simply measured, is a question I leave open here.

What is not an open question are the stakes. Synthetic artifacts are already entering the analytical record, some in bad faith and many in good faith but merely undisclosed. If they harden into the ground truth on which the next AI system trains, the drift will become very hard to reverse.

Competing Interests

The integrity-gating approach described in §4.2 is based on an active research project at the Rochester Institute of Technology led by the author. Additionally, the author may seek investment to support it. Some investors potentially interested in that work may also have an interest in TripleID, whose hardware-based approach is assessed in §4.3. The author currently has no financial relationship with TripleID.

References

1. Plato. *Sophist*.
2. Baudrillard, Jean. *Simulacres et simulation*. Paris: Galilée, 1981.

3. European Data Protection Supervisor. “Synthetic Data.” https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en
4. Searle, John R. *The Construction of Social Reality*. New York: Free Press, 1995.
5. Gettier, Edmund L. “Is Justified True Belief Knowledge?” *Analysis* 23, no. 6 (1963): 121-123. https://courses.physics.illinois.edu/phys419/sp2021/Gettier.pdf
6. Plantinga, Alvin. *Warrant and Proper Function*. New York: Oxford University Press, 1993.
7. ------. “Précis of *Warrant: The Current Debate* and *Warrant and Proper Function*.” Philosophy and Phenomenological Research 55, no. 2 (1995): 393-396. https://andrewmbailey.com/ap/Precis_Warrant.pdf
8. Schwartz, Christopher, and Adam Arthur. “Deceiving the Machine: The Case for Synthetic Evidence as a Cybersecurity Category.” Under review, 2026.
9. National Disease Registration Service, NHS England. *The Simulacrum* (synthetic cancer dataset). https://digital.nhs.uk/ndrs/data/data-outputs/cancer-publications-and-tools/simulacrum
10. Martinez, Antonio Lopo. “Synthetic Evidence: Documentary Deepfakes and the Future of Truth in Brazilian Legal Proceedings.” *Revista Eletrônica de Direito Processual* 27, n. 2 (2026). https://doi.org/10.2139/ssrn.5479486
11. Durand, Maxime. “When Evidence Becomes Synthetic: Admissibility, Authentication, and the Legal Crisis of AI-Generated Proof.” *LexAI Journal*, January 12, 2026. https://lexai.sa.utoronto.ca/when-evidence-becomes-synthetic-admissibility-authentication-and-the-legal-crisis-of-ai-generated-proof/
12. European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). *Official Journal of the European Union* 119, 2016.
13. Bartholdy, Matthias. “Positioning Synthetic Data under EU Data Protection Law.” *Computer Law & Security Review* 61 (2026): 106310. https://doi.org/10.1016/j.clsr.2026.106310
14. Tari, Henry and Adriana Iamnitchi. “Measuring Privacy vs. Fidelity in Synthetic Social Media Datasets.” arxiv.org/abs/2603.03906
15. European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act). *Official Journal of the European Union*, 2024.
16. National Institute of Standards and Technology. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*. NIST AI 100-4. Gaithersburg, MD: NIST, 2024. https://doi.org/10.6028/NIST.AI.100-4
17. California. AI Transparency Act, S.B. 942 (2024), as amended by A.B. 853 (2025).
18. California. Generative Artificial Intelligence: Training Data Transparency, A.B. 2013 (2024).
19. CEN-CENELEC Joint Technical Committee 21, Artificial Intelligence, under European Commission standardisation request M/593 (2023).
20. Coalition for Content Provenance and Authenticity. https://spec.c2pa.org/specifications/specifications/2.4/index.html
21. GB 45438-2025, *Cybersecurity Technology – Labeling Method for Content Generated by Artificial Intelligence* (*网络安全技术 人工智能生成合成内容标识方法*). Mandatory national

standard. Standardization Administration of China, released 14 March 2025, effective 1 September 2025. https://www.codeofchina.com/standard/GB45438-2025.html
22. Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration. *Measures for Labeling AI-Generated Synthetic Content* (*人工智能生成合成内容标识办法*). Issued 14 March 2025, effective 1 September 2025. English translation: China Law Translate. https://www.chinalawtranslate.com/en/ai-labeling/
23. Schwartz, Christopher, Justin Pelletier, David I. Schwartz, Matthew Wright, and Andrea Hickerson. "Deepfakes in Narrative Warfare." In *Artificial Intelligence and International Security*. Eds. Alena Vysotskaya Guedes Vieira, Arshin Adib-Moghaddam and Mohammad Eslami. Manchester University Press, 2026. https://manchesteruniversitypress.co.uk/9781526196163/
24. National Science Foundation. Cybersecurity Innovation for Cyberinfrastructure (CICI). https://www.nsf.gov/funding/opportunities/cici-cybersecurity-innovation-cyberinfrastructure
25. National Institute of Standards and Technology. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST AI 100-2e2025. Gaithersburg, MD: NIST, 2025. https://doi.org/10.6028/NIST.AI.100-2e2025
26. Federal Communications Commission. *Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts*. Declaratory Ruling, CG Docket No. 23-362, FCC 24-17. Adopted 2 February 2024, released 8 February 2024. 39 FCC Rcd 1783. https://docs.fcc.gov/public/attachments/FCC-24-17A1.pdf
27. The DOI Foundation. https://www.doi.org
28. Telephone Consumer Protection Act of 1991, Pub. L. No. 102-243, 105 Stat. 2394 (codified at 47 U.S.C. § 227).
29. NIST Collaborative Research Cycle / SDNist. National Institute of Standards and Technology. https://pages.nist.gov/privacy_collaborative_research_cycle/
30. DataCebo. SDMetrics. https://docs.sdv.dev/sdmetrics/
31. Offenhuber, Dietmar. "Shapes and Frictions of Synthetic Data." *Big Data & Society* 11, no. 2 (2024). https://doi.org/10.1177/20539517241249390
32. Bordes, Florian, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. "IntPhys 2: Benchmarking Intuitive Physics Understanding in Complex Synthetic Environments." arXiv:2506.09849 (2025). https://arxiv.org/abs/2506.09849
33. Bansal, Hritik, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. "VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation." In *The Fourteenth International Conference on Learning Representation*s (ICLR 2026), 2026. https://videophy2.github.io