

The Counterfeit Observation

Synthetic Evidence, Industrialized Spectacle, and the Destruction of
Observational Trust

Flyxion

Independent Researcher

October 2026

Abstract

Mass-produced synthetic media does not merely lower the quality of online entertainment. When it is presented in the manner of documentary recording, it counterfeits the relationship between a representation and the event it claims to preserve. This essay argues that three substitutions reinforce one another. Observation is replaced by fabrication, because what looks like evidence of the world is generated without any corresponding worldly event. Discovery is replaced by engagement selection, because what is published is determined by anticipated audience reaction rather than by the significance of something actually found. Human judgment is replaced by platform priority, because what a user has explicitly refused remains eligible for delivery. The essay distinguishes generative technology from the institutions and incentives that govern its deployment, and it directs its objection at the industrial production of counterfeit observations and at the distribution systems that make them profitable, not at computational assistance, fiction, fantasy, or automation as such. Four mathematical appendices formalize the causal, selective, economic, and constraint-theoretic components of the argument.

1 The World Was Worth Watching

1.1 The original value of recorded experience

Around the year 2000, a university researcher in developmental psychology, linguistics, or animal behavior could spend an entire working day transcribing a few minutes of recorded conversation, or could spend a season watching a single animal population in the hope of documenting one unusual behavior. The labor was justified because the recording held something the investigator could not have invented. The subject did what the subject did. A toddler produced a morphological error that no grammar had predicted, a primate used a tool in a way the literature had not described, and the transcript or field note preserved the surprise so that others could examine it later.

The methods of the period reflect this dependence on an external source of constraint. Field notes were dated and located. Audio recordings were archived with metadata about the setting and the participants. Behavioral coding schemes were tested for inter-rater reliability so that two observers would see the same event in the same footage. Conversation analysis, in the tradition of Sacks, Schegloff, and Jefferson, treated the recorded turn as the primary datum precisely because a transcript of what people actually said exposes regularities that introspection and invention do not supply [9]. Ethnography accumulated thick description because a particular place and a particular afternoon had resisted the ethnographer's expectations. Experimental observation, for its part, was valued in proportion to its capacity to come out otherwise than predicted.

The common feature across these practices is that a recording preserves an encounter with a system that exists independently of the person who records it. A child's unexpected utterance, a gorilla's unusual gesture, or a person improvising an animal rescue with whatever tools are at hand can reveal something that was not already contained in the observer's expectations. The proposition on which the rest of this essay depends can be stated plainly.

Proposition 1 (Provenance condition). *The informational value of observational media depends on its causal relationship to an event that occurred independently of its representation.*

This is not a claim that every recording is true. Recordings can be staged, cropped, mislabeled, and edited. It is a claim about what makes a recording a candidate for evidence in the first place, namely that its production passed through the event it purports

to show.

1.2 Entertainment as epistemic encounter

A substantial category of entertainment derives its value from what it teaches about the world, and only secondarily from the pleasure of stimulation. A surprising animal behavior entertains because it enlarges our sense of what animals can do. An unusual physical event entertains because it challenges expectations about mechanics. A spontaneous act of compassion toward a stranger or an injured creature entertains in part because it demonstrates a possibility of human conduct that the viewer might not have credited. In each case the pleasure is inseparable from the update the viewer performs.

It helps to give this a name. Call *epistemic surprise* an unexpected observation that requires a revision of one's understanding of an independent system. Epistemic surprise differs from mere novelty. A novel arrangement of pixels can be produced without limit and carries no obligation to revise anything. An epistemic surprise obliges the viewer to ask how the world could be arranged such that this occurred, and the answer is informative because the world is not under the control of the person asking.

None of this disparages imaginative entertainment. Fiction, comedy, music, animation, and fantasy have no need to document reality in order to possess value, and nothing in this essay assumes otherwise. A viewer who watches a cartoon squirrel rescue a cartoon child is not misled about squirrels. The difficulty arises in a narrower and more specific case: media that promises discovery and supplies fabricated observation instead. The promise is carried by form, by the handheld framing, the apparent surveillance camera, the caption that announces a real rescue, the audio of a startled bystander. These are the conventions by which an audience learns that it is looking at a recording of something that happened. When the conventions are retained and the event is removed, the viewer is offered the epistemic reward without the epistemic substance.

It helps to separate three kinds of value that a piece of media can carry. Observational value belongs to a recording that reveals something about an independent system, as a genuine gorilla imitation would reveal something about an animal. Artistic value belongs to a work in which a maker has developed an imaginative, comic, or interpretive idea, as a deliberately fictional gorilla performance might. Engagement value belongs to anything that attracts attention, whatever it contains. A video selected only because it holds attention may have engagement value without possessing either of the first two.

The difficulty arises when engagement value becomes the dominant criterion for what circulates, and the three categories are presented in one undifferentiated stream. Writing the human value of a work in a deliberately rough decomposition,

$$U_{\text{human}}(V) = I(V; W) + A(V) + C(V),$$

where $I(V; W)$ is the information the video carries about the world, $A(V)$ its aesthetic or imaginative value, and $C(V)$ its contribution to understanding or connection between people, makes the point that a platform optimizing a different quantity, for instance

$$J(V) = \alpha T(V) + \beta R(V) + \gamma S(V),$$

with T watch time, R reactions, and S sharing, has no mathematical guarantee of increasing the first. The decomposition is conceptual and is not offered as a measurable universal formula for value. Its use is to show that the objectives are distinct and that they can diverge systematically, especially once generative systems remove the cost of producing superficially compelling material.

2 The Counterfeit Observation

2.1 The squirrel and the changing pole

A video that circulated widely on a social platform purported to show the rescue of a squirrel entangled in a tennis net. The apparent event contains several features that a viewer would reasonably find informative. There is entanglement in a man-made object, which is a recurring hazard for urban wildlife. There is human intervention, with the delicate handling that this requires. There is animal distress, an apparent maternal response, and physical movement through a constrained environment of net, pole, and ground.

Yet the pole changes length. Across successive moments of the clip the vertical post against which the action takes place is not the same height, and the net attaches to it in ways that do not cohere. The scene has no persistent physical geometry.

This matters more than an ordinary visual mistake would. Real footage can be shaky, badly lit, and awkwardly framed, but it cannot contradict itself about the dimensions of a fixed object, because the object was there and the camera recorded it. A representation

that cannot hold a pole constant is a representation that was not constrained by a pole. It borrows the conventions of documentary recording without being accountable to the continuity of the depicted scene. The distinction at stake is between an event constraining its representation and a representation inventing its event. In the first case the recording inherits the event's structure whether or not anyone intended it to; in the second the structure is whatever the generator produced, and its apparent consistency is a statistical accomplishment, not a physical one.

2.2 The gorilla that imitates everything

A second family of videos shows a gorilla in an enclosure supposedly imitating body-building routines, gymnastic maneuvers, and a range of other human activities with fluent precision. If such behavior were genuinely observed, it would invite an extensive program of investigation. Researchers would ask about imitation and social learning in great apes, about the biomechanics of a body so differently proportioned performing movements evolved for another, about the limits of comparative cognition, and about how the animal acquired the repertoire. The footage would be an anomaly demanding explanation, and its value would lie in the explanations it provoked.

A fabricated performance provides no corresponding evidence about gorillas. It sets up the same questions and then dissolves the possibility of answering them, because there is no animal whose behavior could be studied. What the video does reveal is something about the generator and about its audience: that a model can render a plausible ape in a plausible gym, and that viewers will watch it. The distinction worth drawing is between a surprising output of a generative model and a surprising observation of an independent organism. Both can be surprising. Only one is a surprise about the world.

A related displacement concerns moral action. An animal rescue is compelling partly because someone chose to help a vulnerable creature when walking away was available, and the moral significance belongs to the act and not merely to its appearance. A synthetic rescue reproduces the emotional cues of compassion without requiring anyone to exercise it. The animal was never endangered, but neither was it helped, and no rescuer assumed any responsibility. The audience receives a simulation of moral satisfaction disconnected from any corresponding moral event. This does not make fictional compassion worthless, since literature and cinema communicate real moral understanding under an interpretive agreement that announces their imaginative status. The harm

arises when the fabricated event is distributed as an unsolicited documentary encounter, which exploits the viewer’s expectation that something actually happened.

2.3 The counterfeit observation as a formal object

The essay’s principal concept can now be stated. A witness is an entity whose testimony is causally connected to an event, and a recording instrument extends witnessing by preserving traces of physical occurrences. In the ordinary observational process an event W in the world gives rise to an observation O , which is preserved as a recording R :

$$W \longrightarrow O \longrightarrow R.$$

Synthetic generation has a different structure. A generative model M , together with a prompt and some source of randomness, produces a recording-like artifact R' directly:

$$M \longrightarrow R'.$$

The resemblance of R' to R does not establish the existence of a purported event W' . This is the sense in which R' is a *counterfeit* observation: it has the sensory form of the output of the first process and the causal ancestry of the second.

Visual plausibility is therefore not equivalent to evidentiary provenance. It must also be acknowledged that conventional photography and video can be staged, cropped, misattributed, or manipulated, and that this was true long before generative models. Forgery of recordings is not new. What is new is the scale, the cost, and the removal of the requirement that any scene exist at all. A staged photograph still required a stage. The relevant change is that the minimal fabrication, one that involves no real location, no real subject, and no real camera, has become as cheap as the minimal truthful recording, and in some respects cheaper.

3 The Industrialization of Fabricated Experience

3.1 The synthetic engagement factory as a production system

The analysis so far has treated individual artifacts. The more consequential unit is the account, the production operation, and the distribution network. Across some platforms

there is an apparent pattern of pages that publish large numbers of sensational animal rescues, impossible athletic performances, sentimental narratives of reunion and sacrifice, and supernatural spectacles, at a cadence no recording practice could sustain and with a uniformity of style that suggests a shared production pipeline.

The concern is not merely that individual videos are synthetic. It is that entire accounts may operate as industrial systems for generating and selecting content with little commitment to documenting or explaining anything. The account is a machine whose inputs are prompts and engagement statistics and whose output is a stream of artifacts tuned to those statistics.

Care is needed in how this is stated. Automation and publication frequency alone are insufficient to establish deception. A prolific human artist may publish daily. An automated scientific observatory may publish thousands of images of the sky, each a faithful record of what its instruments received. These are not counterfeits, because their outputs remain accountable to something outside the production process. The operative difference between such cases and a synthetic engagement factory is not volume or automation but the relationship between what is published and what occurred. The observatory is constrained by the sky. The factory is constrained by the feed.

3.2 Selection without discovery

The mechanism by which a factory becomes effective can be given a simple formal description. Let G_θ be a generator that produces candidate videos $V_1, V_2, \dots$, and let $E(V)$ denote a measure of the engagement a video attracts, such as watch time, shares, or completion rate. A production operation retains or promotes videos according to a selection rule with threshold τ :

$$\mathcal{P}_\tau = \{ V_i \sim G_\theta : E(V_i) \geq \tau \}.$$

Repeated application of such a rule, with the generator updated toward its own survivors, increases the prevalence of content that triggers attention. It does not thereby increase the informational value of the content, because nothing in the rule refers to information. The relationship to well-known statistical hazards is direct. Retaining the maximum of many draws is a multiple-comparison effect, in which the most extreme of a large number of candidates is extreme in part because there were many candidates [1]. Using engagement as a stand-in for value invites Goodhart’s law, according to which a measure

that becomes a target ceases to be a good measure [6, 11, 7]. The proxy and the quantity it was meant to track separate under optimization pressure, and they separate in the direction that the proxy rewards.

There is a further asymmetry in what the audience sees. The viewer encounters the selected spectacle without seeing the unsuccessful generations from which it emerged. A strikingly improbable rescue appears as a single event, when it is in fact the survivor of a large implicit sample. A viewer who correctly reasons that such an event is rare in the world may nonetheless find it common in the feed, and the discrepancy is entirely a consequence of selection.

The distribution the audience samples is therefore not a distribution over events in the world, nor even over imaginable events, but over events conditioned on audience responsiveness, which may be written $\mathbb{P}(V \mid E \geq \tau)$. With repeated generation and selection the producer can refine this distribution toward increasingly effective attention triggers. If engagement is rewarded without any independent requirement of authenticity, there is no necessary incentive to preserve factual accuracy, biological plausibility, or physical consistency, and consistency can become an economic disadvantage whenever an impossible spectacle outperforms an ordinary event. The process resembles artificial selection, except that the fitness criterion is performance within an attention market and not survival in a physical environment.

3.3 The inversion of curiosity

It is useful to contrast two causal processes. In the first, an event gives rise to an observation, and the observation gives rise to human curiosity:

$$\text{Event} \rightarrow \text{Observation} \rightarrow \text{Human curiosity.}$$

In the second, predicted curiosity determines what will be fabricated, and the fabricated event produces engagement, which in turn updates the prediction:

$$\text{Predicted curiosity} \rightarrow \text{Fabricated event} \rightarrow \text{Engagement.}$$

In the first process curiosity follows discovery. The audience cares about what the world turns out to contain. In the second, anticipated curiosity determines what will exist, and the loop closes on the audience itself. An archive produced this way increasingly

describes the responsiveness of audiences, and it describes the external world only by accident.

This analysis helps explain why certain features recur. Engagement-based selection favors gaudiness, because saturated and exaggerated imagery holds attention. It favors exaggerated emotion, because intense affect is shared. It favors implausible coincidences, because they produce the double take that extends watch time, and it favors impossible physical performances for the same reason. These are not defects that better generators will remove. They are the features that the selection rule rewards, and a better generator will render them more convincingly. The very properties that would make a recording informative about the world, its inconvenient particularity, its failure to resolve neatly, its lack of a climax, are properties that the selection process disfavors.

4 The Demon-Haunted Feed

4.1 From animal fabrication to supernatural evidence

Carl Sagan’s *The Demon-Haunted World* remains the most accessible statement of why a society needs habits of skeptical inquiry, and why the absence of such habits leaves it exposed to confident assertion about invisible agents, concealed conspiracies, and miraculous events [10]. Sagan’s central practical device was the baloney detection kit, a set of questions whose purpose is to ask, of any extraordinary claim, what independent confirmation exists and what alternative explanations have been excluded.

Consider in this light synthetic videos depicting a religious figure battling demons, regenerating an amputated limb, restoring a missing eye, or performing other purported miracles in documentary style. The objection is not to religious art. Painted altar-pieces, stained glass, devotional film, and imaginative illustration of supernatural events are long-standing and legitimate forms, and no viewer of a Renaissance annunciation supposes that a camera was present. The objection is to the presentation of invented events in a style that implies empirical confirmation. A fictional squirrel rescue counterfeits a plausible occurrence. A synthetic miraculous healing counterfeits evidence for a claim that would require extraordinary scientific scrutiny. The second case is more serious because the claim being counterfeited is one for which genuine documentation would be of unusual evidential weight, and so its imitation exploits an unusually large gap between what a viewer would demand and what the viewer is shown.

4.2 Manufacturing the evidence that extraordinary claims require

Several kinds of material must be kept distinct. Testimony is a report by a person who claims to have observed something. Illustration is a depiction that makes no claim to have recorded anything. Simulation is a model-driven rendering that purports to show what would follow from stated assumptions. Independently verifiable evidence is material whose connection to an event can be checked by parties who do not depend on the claimant. Each plays a different role in inquiry, and confusion among them is the usual route by which poor reasoning acquires an air of rigor.

Traditionally an extraordinary claim was accompanied by stories or by the testimony of witnesses, and skeptics asked for something stronger, such as photographs or video. Contemporary generative media can manufacture what looks like exactly the audiovisual documentation that skeptics would otherwise request. The video supplies the appearance of confirmation without supplying the underlying observation. It can reinforce preexisting beliefs, create unwarranted confidence, and blur the empirical status of supernatural claims for viewers who have no practice in distinguishing the forms of material listed above. The practice may be called evidentiary counterfeiting: the manufacture of an artifact that resembles the output of an observational process without the corresponding observation. When a person reports that an amputated limb regrew through supernatural intervention, the appropriate response is to ask for independently verifiable documentation, medical records, credible witnesses, and evidence that the event occurred as described. A generative system can now produce something that looks like the requested documentation without producing any of the underlying evidence. The squirrel video counterfeits an ordinary event, the gorilla video counterfeits extraordinary animal behavior, and the miraculous healing counterfeits evidence for a claim that conflicts with well-established biological expectations. The stakes rise along this progression, because the last does not merely add a false event to the record. It can encourage viewers to revise their understanding of what is physically possible on the basis of something that never happened, and the change in the economics of credulity is considerable: extraordinary claims, which once circulated without adequate evidence, can now be accompanied by unlimited quantities of fabricated evidence.

The argument should not be overstated. It does not assume that every viewer is deceived, that most viewers take such material at face value, or that every religious depiction is offered as documentary proof. Many are plainly offered as devotional imagery,

and many viewers know what they are watching. The relevant issue is the misleading evidentiary claim that arises when documentary presentation is combined with fabricated content, and the degree to which a given artifact does this is a matter for case-by-case examination.

4.3 The democratization of counterfeit proof

The mechanism is not confined to religious miracles. The same pipeline can produce apparent evidence of impossible medical recoveries, invented historical events, extraordinary archaeological discoveries, fabricated scientific experiments, and nonexistent natural phenomena. Each of these has a legitimate counterpart. Scientific simulation is a respectable practice, clearly identified fiction is a respectable art, and reconstruction of historical scenes for museums is a respectable educational tool. The difference lies in the representational claim and in whether it is honestly signaled.

Proposition 2 (Abundance without access). *When evidence-like artifacts become inexpensive to manufacture, the apparent abundance of evidence can increase while the availability of trustworthy observations remains unchanged.*

The proposition is almost a tautology once stated, yet its consequences are underappreciated. A cultural environment that equates having many videos of an event with having good evidence of that event has built a rule of inference that now has a cheap exploit.

5 The Collapse of Observational Trust

5.1 When everything remarkable becomes suspicious

The second-order consequences may be more damaging than the first-order deception. It is not only that fabricated events may be believed. Authentic events may also be dismissed because their appearance resembles common synthetic fabrications. A genuine and unusual animal behavior, a spontaneous act of kindness, or a remarkable physical phenomenon may lose credibility simply because it looks extraordinary, and looking extraordinary is now the signature of the fabricated.

A Bayesian treatment makes the structure explicit. Let R denote the proposition that the depicted event occurred, and V the proposition that a video of the described

kind has been encountered. Then

$$\mathbb{P}(R | V) = \frac{\mathbb{P}(V | R)\mathbb{P}(R)}{\mathbb{P}(V | R)\mathbb{P}(R) + \mathbb{P}(V | \neg R)\mathbb{P}(\neg R)}.$$

In odds form,

$$\frac{\mathbb{P}(R | V)}{\mathbb{P}(\neg R | V)} = \underbrace{\frac{\mathbb{P}(V | R)}{\mathbb{P}(V | \neg R)}}_{\text{likelihood ratio}} \cdot \frac{\mathbb{P}(R)}{\mathbb{P}(\neg R)}.$$

The evidentiary value of the video is carried by the likelihood ratio. A video is strong evidence for an event to the extent that videos of that kind would be much more likely to exist if the event occurred than if it did not. When fabrication is rare, $\mathbb{P}(V | \neg R)$ is small and the ratio is large. As synthetic production becomes cheap, $\mathbb{P}(V | \neg R)$ rises toward $\mathbb{P}(V | R)$ for precisely the class of striking, unusual footage, and the likelihood ratio approaches one. The video then barely moves the posterior away from the prior.

Three dependencies deserve emphasis. The first is the prior $\mathbb{P}(R)$, which for extraordinary events is low and which therefore makes the posterior sensitive to any increase in the rate of fabrication. The second is provenance, since information about the source, the chain of custody, and the capture device can restore a large likelihood ratio by conditioning on facts that are hard to counterfeit. The third is the reference class. The relevant $\mathbb{P}(V | \neg R)$ is not the rate of fabrication across all video but across videos of this kind, in this style, circulated through this channel. A viewer who correctly reasons within the right reference class will assign low weight to an unauthenticated viral rescue clip, and will be right to, but will assign the same low weight to a genuine one.

The same structure can be seen from the side of rarity. An unusual event in an archive of authentic recordings carries information precisely because it occurred despite the constraints that reality imposes, and its rarity is part of what makes it interesting. When the same feed also contains an enormous number of synthetic extraordinary events, generated specifically because they are extraordinary, the relationship between rarity and informational value collapses. The spectacular is no longer evidence of an unusual occurrence. It is evidence that a content-generation system has identified something likely to attract attention. The extraordinary becomes commonplace in representation while remaining extraordinary in reality. The resulting harm is better stated as a loss of knowability than as a gain in falsehood. The fundamental damage of synthetic documentary material is not that it makes imaginary things look real, but that it makes real things harder to know, so that people may become less willing to believe truths as

well as more willing to believe falsehoods.

5.2 The transfer of verification costs

A producer of synthetic content can generate a new spectacle at negligible marginal cost, while each viewer who wishes to know whether it is authentic must spend time and attention investigating. The burden of verification shifts from the publisher to the audience, and from the maker of a claim to the people trying to evaluate it. Because the number of viewers is large, the aggregate cost of verification can exceed the cost of production by orders of magnitude, and it is paid by people who did nothing to create the problem.

This asymmetry among production cost, verification cost, and the collective degradation of trust has consequences in every domain that depends on shared observation. Scientific communication suffers when instrumental data, microscopy images, and field recordings must be accompanied by ever more elaborate demonstrations of origin. Journalism suffers when the evidentiary weight of eyewitness video falls and when real footage is dismissed as fabricated, a pattern sometimes called the liar’s dividend [3]. Education suffers when students cannot rely on the visual record as a starting point for inquiry. Archival practice suffers when the volume of plausible but unreliable material overwhelms the capacity to catalog and authenticate. Ordinary social observation suffers as well, because people’s habit of telling one another about something remarkable they saw depends on a background expectation that such reports are usually true.

5.3 What viewers learn from video

Much informal education proceeds through observation. People watch others repair machinery, handle animals, build structures, cook, navigate difficult situations, and carry out experiments, and even when they never attempt these activities themselves the watching contributes to their sense of what the world permits and how people behave within it. Video has become an enormous distributed repository of practical and observational knowledge, and synthetic material presented as documentary footage contaminates that repository. Unlike an openly fictional film, it does not announce that its relationship to reality is imaginative.

Two distinct mechanisms of contamination can be separated. The first concerns

intuition. Expectations about gravity, balance, momentum, the behavior of materials, and the habits of animals are partly built from repeated observation, and an object whose length changes between frames, or an animal whose movements no skeleton could perform, models a world in which those regularities do not hold. A viewer who cannot tell that the footage was unconstrained by physical circumstances may absorb false regularities without ever forming a belief that could be examined and corrected. Children are the most exposed group, since their intuitions are still forming and they are the least likely to ask whether a clip is real. It would be a mistake to state more than this. Direct experience, schooling, and authentic recordings continue to teach physical intuition, and there is at present no measurement of how much synthetic video alters what people learn from the rest. The claim that a generation might grow up with distorted physical intuitions is a risk that follows from a plausible mechanism, and it should be stated as a risk and treated as an empirical question.

The second mechanism concerns the copying of genuine material. Consider a recording of a bat freed from a barbed wire fence by someone who holds the animal in a folded cloth while working the wing loose. The recording is informative because the technique is sound: the cloth protects the person from a bite and the animal from handling that would injure it. Once such a recording attracts attention, it becomes a template. A generator can produce it again from many angles, with many settings and many different people, at negligible cost. Nothing in the generation process is accountable to the correctness of the procedure, so a copy may omit the cloth, show an unsafe grip, or leave out any suggestion of contacting a wildlife rehabilitator. Viewers cannot tell which version is right, and unlike the false physics of an impossible pole, the content of such a video is something people act upon.

The selection model of the preceding sections already describes what follows. Let $c(V) \in \{0, 1\}$ indicate whether a candidate shows a correct procedure, and let p be the fraction of generated candidates for which $c = 1$. If engagement were independent of correctness, the fraction of correct items among the survivors of selection would remain p , which is already no better than the generator’s accuracy. If the incorrect variants are on average more dramatic, for instance because they show a rougher struggle or a more precarious moment, then

$$\mathbb{P}(c = 1 \mid E \geq \tau) < \mathbb{P}(c = 1),$$

and the survivors of selection are systematically less correct than the population from

which they were drawn. Genuine items serve as the seed distribution, and the selection step then drifts the successors away from correctness without anyone intending that result. The original is also displaced. Its maker located the animal, intervened, and recorded the outcome, while the copies cost almost nothing, so the copies can outnumber and outreach the original and take the audience that it earned. This can be called derivative flooding. It dilutes the informational value of the source, and it means that the more useful a genuine recording is, the more likely it is to be imitated in a form that is worse.

The extent to which this occurs, and how often it yields harmful instructions, is not established here. The mechanism is available to anyone with a generator, and that availability, together with the absence of any check on correctness, is sufficient to make it a real hazard.

5.4 Appeals for help and recognition

The preceding sections concern what people learn from recordings. A further domain concerns what people do for one another on the strength of them. Attention is the means by which strangers learn that someone needs help and that someone has done something worth noticing. A household that has lost its home to a fire, a family raising money for a child's medical treatment, and a young person presenting a craft or a science fair project are all making a claim on the attention of people they do not know, and in each case the claim succeeds only to the extent that viewers trust it to be real. The medium that made such claims audible to large numbers of strangers is the same medium now being filled with synthetic imitations of them.

Three difficulties follow, and they differ in kind. The first is competition with optimized imitations. A genuine appeal is bounded by what actually happened to an actual person. Its circumstances are particular and often untidy, and the person making it may photograph badly, speak awkwardly, or be too distressed to present a case well. A synthetic appeal is bounded by nothing of this kind. It can be made more affecting, more photogenic, and more shareable than any genuine case, and a selection process that rewards engagement will favor it for exactly those reasons. The people whose need is real then compete for attention against material tuned to win it, and they compete under a handicap, since the tuning that wins is not available to them without fabricating their own situation.

The second is the discounting of real appeals. Fraudulent appeals for help are older than generative media, but they were costly to produce convincingly, and that cost limited how many could be circulated. When the cost falls to negligible levels, the proportion of unverifiable appeals that are fraudulent can rise, and the reasoning of the Bayesian section applies directly. Write s for the proportion of appeals in an identifiable class that are genuine, and let a prospective helper act only if s exceeds a threshold determined by the cost of being deceived relative to the benefit of helping. When s falls below that threshold, a helper who cannot verify anything declines to help every appeal in the class, and the genuine ones are refused along with the rest. The burden of proof then falls on the person with the real emergency, who must supply documentation, contact details, and corroboration before anyone will act, and who is typically the person least able to do so at that moment. A disaster in which a person must first demonstrate that the disaster occurred is a cost imposed by the environment, and it is a cost that the producers of counterfeit appeals never pay.

The third concerns recognition. A child's craft matters to its maker and to the people near the child partly because it is the child's, made at a particular level of skill with particular imperfections, and the social reward for presenting it is a reward for the effort. A synthetic equivalent can show a more attractive child and a more accomplished object at no effort at all. Its effect is not only to pull attention away from real work, but to establish an apparent standard against which real work appears poor. Where recognition is a significant motive for making things, the existence of an unlimited supply of costless and flawless imitations reduces the return to making the real thing.

It would overstate the case to conclude that people will no longer need to look at one another. Friends, families, and local communities continue to attend to each other through channels that do not depend on public feeds, and nothing in the mechanism described here weakens those relationships directly. The injury is narrower and still serious. The public channels through which strangers learn of one another's needs and achievements work only while their signals are trusted, and mutual aid among strangers depends on such signals. A feed in which genuine appeals cannot be distinguished from manufactured ones degrades the capacity of a society to respond to those who are in trouble and to notice those who have done something good, and it does so while appearing, from the inside, to be full of both.

6 The Misclassification of Human Creativity

6.1 The Sphere Pop illustration

A case that illustrates the limits of labeling concerns an illustration associated with the Sphere Pop project. The work was developed through original pencil drawing, conceptual worked examples, historical source material on the arithmetic traditions it draws upon, and subsequent digital transformation of the drawn material. The intellectual and artistic provenance of the piece was substantially human, and each stage of its development is documented.

Despite this, the platform displayed an AI-related label on the piece. At the same time, synthetic dancing-cat videos encountered elsewhere on the same platform carried no comparable visible label. It would be a mistake to infer the precise reason for either classification without evidence from the platform. The labeling may have been triggered by a metadata field, an editing tool, an automated classifier, or a policy applied unevenly, and observation from the outside does not distinguish these. What the pair of cases does illustrate is the limitation of using a generic production-method label as a measure of authenticity. A label that is attached to a carefully made human work and absent from an industrial synthetic stream is not tracking the property that readers take it to track.

6.2 Human authorship is not a guarantee of truth

The opposite error is equally available, and the essay’s argument should not rest on it. It is tempting to treat material produced without AI assistance as intrinsically more trustworthy. Human authors make mathematical mistakes, misrepresent evidence, fabricate anecdotes, overlook contradictions in their own arguments, and repeat errors that have been in circulation for decades. Computational assistance, for its part, can improve the accuracy of a work when it is used to check independently specified constraints and when its results are appropriately validated, for instance by running a proof through a formal checker or recomputing a table of figures from source data.

The relevant distinction is between AI-assisted verification, in which a tool tests a claim against a criterion that exists outside the tool, and unverified model-generated assurance, in which a model states that something is correct. The former adds a check.

The latter adds a sentence. The logical situation can be recorded in two non-implications:

$$\text{Human authorship} \not\Rightarrow \text{Factual correctness}, \quad \text{AI assistance} \not\Rightarrow \text{Fabrication}.$$

Provenance, authorship, accuracy, and verification are four distinct properties. A work can be human-authored and false, machine-assisted and true, human-authored and verified, or machine-generated and fabricated, and any attempt to collapse these into a single label discards information that readers need.

The point can be pressed further. If the quantity of interest is correctness, there is no general reason to assume that

$$\mathbb{P}(\text{correct} \mid H) > \mathbb{P}(\text{correct} \mid H, A, V),$$

where H denotes human authorship, A AI assistance, and V a successful independent verification process. When the verification is effective and its corrections are themselves properly evaluated, the opposite inequality is often to be expected. A person can spend years on a theory and still make a transcription error, omit a condition from a theorem, miscalculate an example, or introduce an inconsistency into a diagram, and an additional computational review can catch errors that familiarity has hidden from the author. Work produced without any such review has had no independent scrutiny at all, and its freedom from machine assistance is no evidence of accuracy. The qualification is that the involvement of a model is not itself fact-checking. A model may confidently endorse a false proposition or introduce an error of its own, and what matters is whether the procedure supplies additional and preferably independent evidence. A symbolic verifier testing explicit invariants provides stronger assurance than a language model declaring that a proof looks correct.

Two production processes that may both involve artificial intelligence therefore differ entirely in their relation to knowledge. In the first, a human conception passes through formal construction and computational verification to a revised artifact, so that the computation is subordinate to an independently specified set of constraints, such as the rules a mathematical illustration must obey whether or not the result is visually appealing. In the second, an engagement target passes through synthetic generation and audience selection to a published spectacle, and the only constraints are those of audience response. The epistemic value of computation depends partly on whether it is accountable to constraints that exist independently of its own output. This suggests

that a label ought to describe the kind of epistemic work performed. A label reading only that something was made with AI says almost nothing, whereas a statement that a work is human-authored and was checked computationally against a published specification tells the reader something about its reliability.

6.3 From artifact labels to publisher accountability

If artifact-level labels are poor proxies, a more informative alternative is transparency about publishing operations. Relevant account-level characteristics might include automated publication patterns, the repeated presentation of synthetic events as documentary footage, coordinated duplication across many accounts, misleading claims about provenance, and persistent violations of the platform’s own authenticity policies. These are properties of a publisher’s conduct over time, and they are observable in a way that the question of whether a particular image passed through particular software is not.

The distinction may be summarized as one between artifact provenance and publisher accountability. A disclosure about a tool says something about how an artifact was made. A disclosure about a production operation says something about the incentives, accountability, and reliability of the entity that repeatedly presents artifacts to a reader, and it is the second that usually determines whether that entity belongs in a person’s information environment at all. A page that publishes a mathematical illustration every few weeks, with explanation, examples, and references to a continuing body of work, is not a content factory even if its author uses AI extensively for drafting, verification, typesetting, and illustration. A page that publishes dozens of purported animal rescues and miraculous discoveries every day, with repeated violations of physical continuity, no verifiable provenance, and nearly identical engagement formulas, warrants scrutiny even if some individual items were produced by hand. A content factory could be entirely human-operated, and an automated scientific observatory could publish thousands of valuable records daily, so the relevant variable is organizational and epistemic and not technological. A more informative label than a generic mention of AI might read, for example, mass-produced synthetic entertainment account, accompanied by the evidence for the classification and an effective means of excluding such accounts altogether. Such a label would describe the information environment without stigmatizing the tools of individual artists.

Any such regime has risks that must be taken seriously. False positives would harm

legitimate publishers. Pseudonymous publishing is valuable and should not be penalized merely because it is pseudonymous. Automated accounts that serve scientific or educational purposes, such as an account that posts daily satellite imagery or the output of a seismograph, would be swept up by crude volume heuristics and must be protected by criteria that attend to the relationship between content and source. The central question is not whether an individual image has passed through AI software. It is whether an ongoing publisher is behaving as an accountable source of information.

7 Refusal, Agency, and Platform Power

7.1 The unwanted Reel

A common experience on short-video platforms is the repeated rejection of a format followed by its reappearance. A user marks a clip as not interesting, blocks the account that posted it, mutes its audio, and nonetheless encounters similar clips from new accounts, or finds that audio preferences do not persist from one session to the next. It is important to separate what can be observed from claims about intentions. The observable experience is the failure of expressed refusals to bind subsequent delivery. Claims about what particular employees intended are not warranted by anything available from the outside and are not needed for the argument.

A plausible structural explanation lies in how recommendation systems treat attention. If watch time is read as a positive signal, then a user who watches a clip in order to investigate it, to criticize it, or to confirm that it is as bad as expected has produced a signal that the system interprets as interest. The user's negative judgment is invisible at the level of the signal. A user who has gone further and explicitly blocked a format has expressed a considerably stronger preference than mere non-engagement, yet the system may continue to treat the format as eligible because the blocking applies to a specific account and the format is produced by many.

The measurement problem is easily illustrated. Consider a viewer who watches three minutes of an authentic video of unusual animal behavior, because the viewer is surprised and trying to understand what happened. The platform records three minutes of attention and perhaps a replay. Consider a second viewer who watches three minutes of a synthetic video, replays the moment where a pole changes length, comments that the video is fabricated, and shares it with a friend to show how absurd it is. From the

standpoint of a system optimized around observable engagement the two interactions are nearly indistinguishable, and the second video may even score higher. The platform can measure that attention occurred without distinguishing whether it represented learning, delight, skepticism, irritation, or disgust. This is a standard proxy problem: what is optimized is not what made the activity worthwhile in the first place.

7.2 The asymmetry of consent

The distinction at issue is between recommendation and imposition. A recommendation is an offer that the person remains free to decline, and declining has a lasting effect. An imposition is delivery that proceeds regardless of the person's expressed wishes. A platform can be highly successful at identifying content that attracts attention while failing to respect a user's explicit refusal to encounter that content, and the two facts are consistent because the optimization objective rewards the first and is indifferent to the second.

A principle follows. A refusal should persist unless deliberately reversed. This is a modest demand. It asks only that a negative instruction have at least the durability of a setting, and it connects to the difference between user agency and platform optimization. The matter is particularly pressing where users remain on a service for the sake of correspondence, communities, or existing audiences, and not for the entertainment feed. For such users, departure carries high costs, and the feed is imposed as a condition of access to the things they actually came for.

This describes a dependency trap. A person wants access to a social network, and the price of access increasingly includes exposure to an entertainment product that was never requested. A system that continues to present a rejected format can read every failure to leave as evidence that the intrusion is acceptable, although the person may be staying for friends, communities, or correspondence unrelated to the feed. The cost of refusal is transferred to the person who refuses, who must keep restoring conditions already selected, as when a muted audio setting must be muted again. Such a system has an asymmetric model of the person using it, because it registers what is accepted and treats what is rejected as negotiable. It is useful to distinguish a recommender that learns what someone wants from a distribution system that continually tests how much unwanted material someone will tolerate. A refusal that must be repeated indefinitely does not function as a refusal, and has been reduced to a temporary obstacle in the

platform's distribution process.

7.3 The social network against the social person

The Sphere Pop page, with its audience of approximately twenty-six thousand followers, offers an instance. It is a potential channel for communicating an independently developed intellectual project, in which readers can encounter diagrams, worked arithmetic, and historical notes. The environment surrounding that channel encourages consumption of industrially generated spectacle, and the same interface that carries the project carries the counterfeit rescue and the synthetic miracle.

It is worth contrasting two things that are commonly given a single name. A social network is an infrastructure for communication among persons who have chosen to be connected. A social platform, in the sense relevant here, is an engine for maximizing content consumption, which uses the network as a source of attention. The first serves the purposes for which users maintain relationships. The second serves the economic success of the operator. The two can coincide, and often have. The question raised by the present moment is what happens when the economic success of the platform becomes increasingly detached from the purposes for which its users maintain relationships there [4].

8 Toward an Epistemically Accountable Media Environment

8.1 Distinguishing representation from evidence

A practical framework begins from the observation that different media make different claims, and that the standard of assessment should follow the claim. An illustration need not document an event, and it is not defective for failing to. A simulation should disclose its assumptions when it is presented as an explanation. A documentary recording should be accountable to its source, meaning that a viewer should be able to learn when, where, and by what means it was captured. A mathematical representation should be evaluated against its formal specification. These standards are familiar from the practice of the disciplines in which each form is native.

Goffman's analysis of frames is useful here [5]. A frame is the set of expectations

that tells participants what kind of activity is going on, whether a game, a rehearsal, a demonstration, or a deception. The harm of counterfeit documentary media can be described as framing fraud: the viewer is led to apply the frame appropriate to recorded observation to an artifact that belongs to another. The relevant standard depends on the representational claim, not merely on the production technology. A synthetic image, honestly framed as such, is as legitimate as a painting, and a camera recording, dishonestly framed, is as illegitimate as a forgery.

8.2 Accountability at the level of production

Several measures merit consideration. Platforms could provide meaningful transparency about high-volume synthetic publishers, support reliable account-level exclusion so that a user’s refusal of a publisher extends to its successors, and distinguish verified observations from unverified documentary-style fabrications. Provenance standards, such as cryptographically signed capture metadata of the kind developed by industry coalitions, offer a technical basis for the third, though they establish where a file came from and not whether what it shows is true [2].

The limitations are real and should be stated alongside the proposals. Automated detection is an arms race in which the defender must be right every time. Appeals are essential because errors will occur. There is a danger in concentrating excessive authority over truth in the hands of platform operators, who have commercial interests and no special competence at adjudicating empirical claims. The aim is not universal certification of reality, which would be neither achievable nor desirable, but a more defensible allocation of evidentiary responsibility, in which those who publish claims of observation bear some of the cost of supporting them.

8.3 Restoring the right not to watch

Persistent negative preferences should be treated as a fundamental feature of user-directed media. The right to reject a format, a publisher, or a class of content should not depend on repeatedly training an engagement algorithm through irritation and dismissal. Each act of irritation consumes the user’s attention and, as argued above, may be recorded as engagement. This is especially important when content categories are numerous, interchangeable, and inexpensive to reproduce, since the cost of blocking by individual account grows without bound while the cost of producing a replacement ac-

count is close to zero. A refusal expressed at the level of format or class, and honored as a constraint, is the only approach that scales against a producer who can generate variants at will.

8.4 The preservation of human witnessing

The essay may conclude by returning to the historical achievement of ubiquitous recording technology. For the first time, vast numbers of ordinary people can document rare animal behaviors, unusual physical events, spontaneous human interactions, and local phenomena that previously might never have entered the observational record. A person with a telephone can capture the flight of an unfamiliar bird, the collapse of a bridge, the repair of a machine, or an act of kindness on a street corner, and can share it with strangers across the world. This is an extraordinary expansion of humanity's capacity to observe itself and its surroundings.

The tragedy of industrial synthetic documentary media is that it can undermine the trust necessary to benefit from that expansion. The goal should not be to eliminate computational creativity, which has produced real art and real insight, but to preserve the distinction between imagining something and discovering that it happened.

The camera was once valuable because the world had to stand before it. The animal had to behave, the child had to speak, the machine had to operate, and the witness had to be present. A recording could be incomplete, misleading, or fraudulent, but its ordinary production still required an encounter with something outside the representation. Industrial synthetic media dissolves that requirement while retaining the appearance of its fulfillment. We are not merely surrounding ourselves with imaginary events. We are constructing an information environment in which the appearance of having witnessed something increasingly substitutes for the act of witnessing it. The danger is not that machines have learned to imagine. It is that institutions have discovered how to monetize the appearance of discovery while making discovery itself harder to recognize.

A Epistemic Value and Causal Provenance

Let W be a random variable representing the state of the world relevant to some question, taking values in a set $\mathcal{W}$ with prior distribution π . An *observational channel* is a conditional distribution $K(r \mid w)$ describing how a recording R is produced given $W = w$. A *synthetic generative channel* is a pair consisting of a model M with a latent state Z and a prompt ρ , so that $R' = g_M(\rho, Z)$ with Z independent of W .

Proposition 3. *For a synthetic channel in which (ρ, Z) is independent of W , the recording R' is independent of W and $I(W; R') = 0$.*

Proof. Since R' is a function of (ρ, Z) , and (ρ, Z) is independent of W , the pair (W, R') is independent. Mutual information between independent variables is zero. $\square$

In contrast, for an observational channel with any nontrivial dependence of R on W , $I(W; R) > 0$. The point of Proposition 3 is not that synthetic media cannot be informative about *anything*. A recording produced by a model is informative about the model, the prompt, and the population of prompts a producer finds profitable. It is uninformative about the particular worldly event that its form implies, and this is the sense in which the counterfeit has the form of an observation while having none of its content.

The claim can be stated in terms of conditional independence. For the synthetic process, $W \perp R' \mid M$, and, conditional on the generator, the recording carries no information about the world. For the observational process, $W \not\perp R$ in general, and the dependence is what supports inference from recordings to events.

Two refinements are needed. The first concerns hybrids. A generative model may be conditioned on real data, as when an image is edited, upscaled, or reconstructed from a partial recording. Then R' depends on W through the conditioning input $C = h(W)$, and $I(W; R') \leq I(W; C)$ by the data processing inequality. The recording can carry no more information about the world than the input from which it was derived, and the generative step adds appearance without adding information. This gives a precise sense to the intuition that enhancement cannot restore detail that was never captured.

The second refinement concerns the viewer. The viewer does not observe the channel but only R or R' , and cannot in general tell which process generated the artifact. Let $B \in \{0, 1\}$ indicate whether the artifact came from the observational channel. The

viewer’s inferential task is to compute a posterior over W that mixes over B :

$$\mathbb{P}(W \mid r) = \mathbb{P}(B=1 \mid r) \mathbb{P}(W \mid r, B=1) + \mathbb{P}(B=0 \mid r) \mathbb{P}(W \mid r, B=0).$$

Since $\mathbb{P}(W \mid r, B=0) = \pi(W)$, the posterior is a convex combination of the observational posterior and the prior, with weight given by the probability that the artifact is genuine. Provenance information is valuable precisely because it moves $\mathbb{P}(B=1 \mid r)$.

Remark 1 (Limits of mutual information). Mutual information is a measure of statistical dependence and not of epistemic value. It does not reflect which aspects of W matter for a decision, it is indifferent to the semantic content of what is communicated, and it can be large for dependence on features of the world that are irrelevant to any question the viewer cares about. A decision-theoretic measure, such as the expected value of sample information with respect to a loss function, is more appropriate when the aim is to assess the usefulness of an observation. The mutual information criterion is used here only to separate the case of zero dependence from the case of positive dependence, for which it is adequate.

B Engagement Selection and Distributional Distortion

Let a generator G produce candidate items $V_1, \dots, V_N$ independently, each with an engagement score $E_i = E(V_i)$ drawn from a distribution F on the real line. Let the selection rule retain item i if $E_i \geq \tau$. The number retained is binomial with parameters N and $q = 1 - F(\tau)$, so the expected number of retained items is Nq .

Selection from a mixed stream

Suppose the platform receives two kinds of items. Genuine observations arrive at rate n_g per period, and each has engagement distribution F_g . Synthetic items are produced at rate n_s and have engagement distribution F_s . Define the tail probabilities $q_g = 1 - F_g(\tau)$ and $q_s = 1 - F_s(\tau)$. The fraction of selected items that are genuine is

$$\phi(\tau) = \frac{n_g q_g}{n_g q_g + n_s q_s}.$$

Proposition 4. *If $q_s > 0$ and $n_s/n_g \rightarrow \infty$ with q_g and q_s fixed, then $\phi(\tau) \rightarrow 0$. If, in addition, the synthetic producer can adapt G so that q_s increases while q_g is fixed by the world, then ϕ decreases further.*

Proof. Rewrite $\phi = (1 + (n_s q_s)/(n_g q_g))^{-1}$. The ratio $(n_s q_s)/(n_g q_g)$ diverges as $n_s/n_g \rightarrow \infty$ when $q_s > 0$, so $\phi \rightarrow 0$. Monotonicity in q_s is immediate from the same expression. $\square$

The proposition states in simple terms what the factory achieves. Genuine observations are produced at a rate set by the world, and their engagement distribution is whatever it happens to be. Synthetic production is limited by compute and cost, and its engagement distribution is under the producer’s control. Increasing volume and tuning the generator both increase the share of spectacular content that is synthetic, and they do so without any increase in the number of observations of the world.

Repeated optimization

Let the producer iterate. At round t the generator G_{θ_t} produces N candidates, the top fraction α by engagement are retained, and θ_{t+1} is chosen to maximize the likelihood of the retained set. Let $\mu_t = \mathbb{E}_{G_{\theta_t}}[E]$ be the mean engagement of the generator. Under standard assumptions of a location family with a log-concave density, the mean of the retained set exceeds μ_t by an amount that depends on α , and the update moves μ_{t+1} toward that mean. The sequence μ_t is then nondecreasing and bounded only by the support of the engagement measure. Meanwhile, the informational utility $U(V)$ of an item with respect to any fixed world question does not enter the update, so $\mathbb{E}[U]$ under G_{θ_t} is not constrained to increase and, by Proposition 3, is zero for every synthetic item with respect to its implied event.

Divergence between engagement and utility

Define the engagement-weighted distribution over items retained by threshold selection as $p_\tau(v) \propto p_G(v) \mathbf{1}[E(v) \geq \tau]$. The divergence of p_τ from the unselected generator distribution is

$$D_{\text{KL}}(p_\tau \parallel p_G) = -\log q.$$

This quantity grows as the threshold becomes more stringent. It measures the amount of information about the engagement criterion that the selection process injects into

what the audience sees, and it says nothing about the relationship of selected items to the world. The audience’s feed is thus a sample from a distribution that has been tilted by $-\log q$ nats toward whatever the criterion rewards. For a threshold set to retain one item in a thousand, the tilt is $\log 1000$ nats, regardless of the content of the items.

Remark 2. The model is deliberately simple. Real recommendation involves personalization, feedback from the audience to the generator, and strategic behavior by producers who anticipate the selection rule, and the work on performative prediction describes how deployed predictors alter the distributions on which they are evaluated [8]. These features strengthen the conclusion of Proposition 4 without altering its form.

C Verification Asymmetry and Trust Degradation

Let a synthetic producer incur cost c_p to generate and publish one item. Let a viewer incur cost c_v to verify one item, where verification means gathering enough information about provenance to move the posterior probability of authenticity above a threshold. Let m be the number of viewers who encounter the item. If every viewer verified independently, the total verification cost would be $m c_v$, and the ratio of social verification cost to production cost is

$$\rho = \frac{m c_v}{c_p}.$$

For viral items, m may be in the millions, while generation cost c_p may be of the order of the cost of a few seconds of computation. Verification need not be performed independently by every viewer, since shared verification by a few specialists can serve many. Even so, the structure is asymmetric, because the producer bears a cost that does not scale with m and the audience bears costs that do, whether directly or through the degradation of the shared signal.

Rational response to the asymmetry

Consider a viewer who must decide whether to verify, to accept, or to dismiss an unauthenticated documentary-style item. Let s denote the viewer’s estimate of the fraction of such items that are authentic, let L be the loss incurred by accepting a counterfeit, let G_a be the gain from accepting an authentic item, and let L_d be the loss incurred by dismissing an authentic item. The expected payoffs of accepting and dismissing without

verification are

$$u_{\text{acc}} = s G_a - (1 - s) L, \quad u_{\text{dis}} = -s L_d.$$

Accepting is preferred when $s(G_a + L_d) > (1 - s)L$, that is, when $s > s^*$ with $s^* = L/(L + G_a + L_d)$. Verification with cost c_v and perfect resolution has payoff $s G_a - c_v$ and is preferred to both unverified options when c_v is lower than the expected loss avoided by verifying. The set of cases in which verifying is rational shrinks as c_v grows and as the stakes shrink, and for the large class of low-stakes entertainment items viewers will not verify.

Proposition 5. *Suppose $s < s^*$ for an identifiable class of unauthenticated items, and suppose the class contains both authentic and synthetic items that viewers cannot distinguish without verification. Then a rational viewer who does not verify dismisses every item in the class, and authentic items in the class receive no more credence than synthetic ones.*

Proof. For $s < s^*$ the inequality above reverses, so $u_{\text{dis}} > u_{\text{acc}}$. The viewer's action depends only on the class and not on any feature distinguishing authentic from synthetic items, because by assumption no such feature is available without verification. $\square$

The consequence for authentic producers is a form of externality. The producer of a genuine observation bears the cost of capture and publication, but the credibility the observation receives is a function of s , which is determined by the behavior of other producers. If the share of synthetic items in the class rises, s falls, and the return to authentic publication falls with it. Where this return is part of what sustains authentic recording, for instance in the form of attention to a naturalist's account or a local journalist's report, a reduction in the number of authentic observations can follow, which reduces s further. The dynamics admit a stable low-trust equilibrium.

The same decision structure governs appeals for help. Let a prospective helper face an unverifiable appeal drawn from a class in which the genuine fraction is s , let G_a be the benefit of helping a genuine case, and let L be the loss from helping a fraudulent one. The helper acts without verification only when $s > s^*$, as above. A genuine appellant in a class with $s < s^*$ receives no help unless they pay a verification cost c_a themselves, by supplying evidence that moves the helper's estimate of authenticity above the threshold. This cost is borne by the person least able to bear it, and it is generated by the behavior of producers of counterfeit appeals who pay nothing comparable. In this setting the externality of the main text takes its most direct form: the cost of low-cost counterfeit

production falls on the people for whom credibility is a condition of receiving assistance at all.

A simple dynamic

Let a_t be the number of authentic items produced at time t and b_t the number of synthetic items, so that $s_t = a_t/(a_t + b_t)$. Suppose authentic production responds to credibility according to $a_{t+1} = a_t(1 + \kappa(s_t - s^*))$ for a response coefficient $\kappa > 0$, while b_t is determined by cost and engagement and is increasing in t . Then for $s_t < s^*$ authentic production is declining, and since b_t is not, s_t is declining. The state $s = 0$ is an absorbing state of the dynamic, and the state $s = s^*$ is unstable. The model is stylized, but it makes the structure of the argument visible: a cheap producer on one side of the market can drive the equilibrium credibility below the level at which the other side finds it worthwhile to participate.

Mitigations enter the model as ways of lowering c_v or of shifting verification to the producer. Signed capture metadata lowers c_v for items that carry it, and it raises s within the verified class even when s is low overall. This is the sense in which provenance infrastructure is a public good, since it restores the separability of authentic and synthetic items on which Proposition 5 depends.

D Negative Preferences and Persistent Refusal

Let $\mathcal{I}$ be a set of items, $\mathcal{U}$ a set of users, and $\mathcal{T}$ a set of discrete times. A recommender at time t for user u selects an item according to a policy $\pi_t(\cdot | u)$ over $\mathcal{I}$. The standard design maximizes predicted engagement,

$$\pi_t(\cdot | u) \in \arg \max_{\pi} \mathbb{E}_{i \sim \pi} [\hat{e}_t(u, i)],$$

where $\hat{e}_t(u, i)$ is the predicted engagement of u with i learned from past interactions.

Preference inference versus binding prohibition

A user may express a refusal through an action such as blocking an account, muting a sound, or marking a format as unwanted. The system may treat such an action in two

different ways. In the first, preference inference, the action is a feature that updates $\hat{e}$, typically by lowering the predicted engagement of similar items. In the second, binding prohibition, the action adds to a constraint set $B_t(u) \subseteq \mathcal{I}$ of items that must not be delivered, and the policy must satisfy $\text{supp } \pi_t(\cdot \mid u) \cap B_t(u) = \emptyset$.

Proposition 6. *Under preference inference with a finite penalty λ , so that the objective for a blocked item i is $\hat{e}_t(u, i) - \lambda$, there exist engagement predictions $\hat{e}_t(u, i)$ for which the policy selects i . Under binding prohibition, no such prediction exists.*

Proof. Under preference inference, item i is selected whenever $\hat{e}_t(u, i) - \lambda$ exceeds the predicted engagement of every unblocked alternative. If $\hat{e}_t(u, i)$ is unbounded above, which holds for optimized spectacle as engagement is trained to increase, there are values for which this holds. Under binding prohibition, the feasible set excludes i by construction, so no value of $\hat{e}_t(u, i)$ can lead to its selection. $\square$

The proposition is a formal statement of an intuition about the relationship of refusal to optimization. A soft penalty is an invitation to compete, and the optimizer will find items that win the competition. A constraint is not a competition.

Persistence

Persistence requires two further conditions. The first is monotonicity of the constraint set: $B_t(u) \subseteq B_{t+1}(u)$ unless the user performs an explicit reversal action at some time s with $t < s \leq t + 1$. The second is that the constraint set is closed under an equivalence relation $\sim$ that identifies items belonging to the same refused class. A refusal of an account should bind its successors and aliases, and a refusal of a format should bind new items of that format. Formally, if $i \in B_t(u)$ and $i \sim j$ then $j \in B_t(u)$.

Definition 1 (Respect for persistent refusal). A recommender respects persistent refusal if for every user u , every time t , and every item i in the closure of the set of refused items under $\sim$, and in the absence of an explicit reversal by u , $\pi_t(i \mid u) = 0$.

The role of the equivalence relation is the point at which a design either scales against cheap variation or does not. If $\sim$ is the identity relation, then a refusal binds only the exact item or account refused, and a producer who can generate variants at negligible cost can evade it indefinitely. If $\sim$ is too coarse, it harms users by excluding wanted items. The choice of $\sim$ is therefore a design decision that should be exposed to the user,

who can inspect and edit the classes they have refused. In the case of account-level exclusion, one natural proposal is that $\sim$ include accounts that share operator identity or coordinated publishing infrastructure, subject to the safeguards on false positives discussed in the main text.

Cost of respect

A recommender that respects persistent refusal incurs a bounded loss in predicted engagement. Let $\mathcal{I}_t(u)$ be the feasible set before applying constraints and let $\mathcal{I}'_t(u) = \mathcal{I}_t(u) \setminus B_t(u)$ be the feasible set after. The loss in maximal predicted engagement is

$$\Delta_t(u) = \max_{i \in \mathcal{I}_t(u)} \hat{e}_t(u, i) - \max_{i \in \mathcal{I}'_t(u)} \hat{e}_t(u, i) \geq 0.$$

The loss is positive only when the highest-engagement item is one the user has refused. In that case, the loss is the cost of honoring a decision that the user has made, and it is a cost that the user, not the platform, is entitled to impose. That the platform objective registers it as a loss is not an argument against the constraint. It is a measurement of how much the objective and the user’s expressed wishes diverge, and the divergence is the phenomenon this appendix is meant to describe.

References

- [1] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1), 289–300.
- [2] Coalition for Content Provenance and Authenticity. Technical specification for content credentials. Published by the C2PA.
- [3] Chesney, R., and Citron, D. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review*, 107, 1753–1819.
- [4] Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
- [5] Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press.

- [6] Goodhart, C. A. E. (1975). Problems of monetary management: the U.K. experience. In *Papers in Monetary Economics*, Volume I. Reserve Bank of Australia.
- [7] Manheim, D., and Garrabrant, S. (2018). Categorizing variants of Goodhart’s law. arXiv:1803.04585.
- [8] Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*.
- [9] Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- [10] Sagan, C. (1995). *The Demon-Haunted World: Science as a Candle in the Dark*. Random House.
- [11] Strathern, M. (1997). ‘Improving ratings’: audit in the British university system. *European Review*, 5(3), 305–321.