Cruelty Across Substrates:

From Animal Cruelty to Embodied AI and the Coming Detection Gap

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, May 2026 (v7.5)

ABSTRACT

Childhood cruelty to animals is among the most durable behavioral warning signs of later interpersonal violence in the forensic and clinical literature. The predictive mechanism does not rest on any settled answer about the moral status of the target. It rests on what the perpetrator practices, rehearses, and trains in themselves: desensitization, dominance affect, and the social learning of cruelty as pleasure. This paper offers a conceptual continuity argument and research agenda. We argue that the substrate of the target is incidental to that training, with one critical qualification: the configuration of the target matters. Routine digital destruction in static media (the static video game target, the scripted non player character) is a categorically different case from the closed loop, contingent emotional response systems on which the argument turns. As world models, embodied robotics, and closed loop emotional voice systems converge toward agents that perform distress in ways a user cannot reliably distinguish from a being in pain, we argue that the dispositional consequences of practiced cruelty against those agents are likely to recruit sufficiently similar mechanisms to warrant clinical and forensic attention. The paper does not claim that synthetic agents are moral patients. It claims that the warning signal forensic and clinical fields have relied upon for fifty years risks migrating to a venue with substantially diminished detection infrastructure. Existing human robot interaction research already shows that users do not need to believe a synthetic agent is conscious to respond to it through social, emotional, or dominance based scripts. The Kantian observation that cruelty corrupts the cruel is sufficient to ground the argument. Conduct disorder assessment, character evidence frameworks, school counseling protocols, and pre employment screening for sensitive roles will need to adapt or lose one of the more durable early behavioral indicators available to forensic and clinical assessment. The paper is offered as agenda setting; the operational protocols it identifies as necessary are deliberately scoped to a planned companion paper.

Keywords: Forensic Psychology, Animal Cruelty, Conduct Disorder, Human Robot Interaction, Embodied AI, Synthetic Cruelty, Detection Infrastructure, AI Ethics

I. INTRODUCTION

A child who tortures the family cat is not, in the eyes of most criminologists or clinical psychologists, simply doing something to the cat. The target matters morally and legally, but the diagnostic significance of the act runs in the other direction. The behavior signals something about the perpetrator. It tracks with later interpersonal violence in ways that have held up across decades of replication, across changes in research methodology, and across shifts in

clinical nomenclature. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, lists physical cruelty to animals as one of the explicit symptom criteria for Conduct Disorder (American Psychiatric Association, 2013). The Federal Bureau of Investigation has incorporated patterns of childhood animal cruelty into behavioral analysis since the foundational work of the Behavioral Science Unit in the late twentieth century (Ressler, Burgess, and Douglas, 1988).

The reason this warning signal works is not because animals are moral patients. The reason it works is that practicing cruelty trains cruelty. The child who derives pleasure from the suffering of another being rehearses a disposition. The disposition generalizes. The literature is dense and the prediction is robust.

This paper makes a single argument. The substrate of the target is incidental to that training. If the perpetrator cannot reliably distinguish the target's distress performance from a being in pain, the disposition is likely trained through sufficiently similar mechanisms regardless of what is behind the lights. Recent and rapidly accelerating developments in world models, embodied artificial intelligence, and closed loop emotional voice synthesis are pushing the substrate question into operational territory. Systems shipped within the past six months can already match a user's emotional state in audio in under two hundred milliseconds (Inworld AI, 2026; MarkTechPost, 2026). Embodied robotics platforms now train on world models built explicitly to learn from environmental experience the way humans and animals learn (World Labs, 2025; Google DeepMind, 2025). The trajectory is not speculative. It is funded, deployed, and accelerating.

The forensic implication is twofold. First, the dispositional training the literature has documented for fifty years is increasingly practicable against synthetic agents whose distress performance is, by design, increasingly indistinguishable from biological distress at the level of interactional uptake, even where users do not attribute consciousness to the system. Second, the social signals that have made animal cruelty a useful forensic indicator (witnesses, veterinary records, school referrals, neighbor complaints, and so on) do not exist in the venue where synthetic cruelty will be practiced. A child screaming at and abusing an embodied AI agent in a bedroom leaves none of the trace evidence that has fed the early warning system for decades.

The paper proceeds as follows. Section II reviews the forensic baseline. Section III articulates the mechanism, what the author calls the Kantian hinge, that licenses the argument without requiring resolution of the consciousness question. Section IV documents the trajectory of current AI development toward humanlike interaction. Section V reviews the human robot interaction evidence on user response to synthetic agents, providing the empirical bridge between Sections II through IV and the convergence claim. Section VI draws the boundary line between routine digital destruction in static media and the closed loop systems on which the argument turns. Section VII states the convergence claim. Section VIII articulates the detection collapse. Section IX addresses forensic, clinical, and policy implications, including explicit privacy and safeguard considerations. Section X responds to the substratism objection. Section XI takes up the animacy awareness question, the most serious philosophical challenge to the argument. Section XII addresses limitations and future work, including the planned companion paper on operational clinical and forensic protocols.

II. The Forensic Baseline

The link between childhood cruelty to animals and later interpersonal violence has been studied since at least Kellert and Felthous (1985), whose interviews with incarcerated offenders established higher prevalence of childhood animal cruelty among aggressive criminals than among non aggressive criminals or non criminal controls. The finding has been replicated across populations, methodologies, and decades. Arluke, Levin, Luke, and Ascione (1999) in the Journal of Interpersonal Violence linked animal abusers to a fivefold increase in violent crime arrests compared with non abusers. Merz Perez and Heide (2004), in the volume Animal Cruelty: Pathway to Violence Against People, synthesized two decades of research into the strongest book length statement of what is sometimes called the graduation hypothesis:

that cruelty against animals in childhood is, for a meaningful subset of perpetrators, a precursor stage to later cruelty against humans.

Wright and Hensley (2003) tested the graduation hypothesis directly against case histories of serial murderers and found a consistent pattern of childhood animal cruelty preceding interpersonal violence in their sample. The behavioral profiling literature initiated by Ressler, Burgess, and Douglas (1988) at the FBI's Behavioral Science Unit had already incorporated the pattern into operational profiling decades earlier. Cleckley's foundational work on psychopathy (Cleckley, 1976; original 1941) and Hare's subsequent development of the Psychopathy Checklist (Hare, 1991) established cruelty in early life as a recurring feature of the antisocial personality profile.

The clinical apparatus formalized the finding. The DSM-5 (American Psychiatric Association, 2013) explicitly names physical cruelty to animals as one of fifteen possible Conduct Disorder symptoms in childhood and adolescence. Conduct Disorder itself is one of the strongest pediatric predictors of adult Antisocial Personality Disorder, which is in turn associated with elevated rates of interpersonal violence (Moffitt, 1993; Robins, 1978). The diagnostic chain is not theoretical. It is built into the document that governs how American psychiatrists and clinical psychologists assess children every day.

Ascione's developmental work (Ascione, 1993; Ascione and Arkow, 1999) added texture to the picture. Children who harm animals are often themselves victims of abuse, witnesses to family violence, or both. The cruelty is not always sadistic in the popular sense. It is sometimes imitative, sometimes a response to trauma, sometimes a means of asserting control in households where the child has none. The diagnostic significance does not disappear in any of these cases. The disposition trained by the practice generalizes regardless of what motivated the practice in the first place.

The case history literature is dense with named examples. Jeffrey Dahmer, Edmund Kemper, Dennis Rader, Albert DeSalvo, Carroll Cole, Ted Bundy, Ian Brady, and others reported childhood animal cruelty as documented antecedents to their adult crimes. The pattern is strong enough that childhood animal cruelty appears in modern violent crime risk assessment instruments and in pre trial behavioral evaluation protocols (Hensley, Tallichet, and Singer, 2006).

The forensic baseline is therefore established. The behavioral pattern is documented. The diagnostic pathway is formalized in clinical nomenclature. The mechanism, however, is what licenses the argument that follows.

III. The Mechanism

Why does childhood animal cruelty predict later interpersonal violence? The answer in the literature is not that animals and humans share moral status, although many of the cited authors hold that view independently. The answer is that the behavior trains a disposition, and the disposition generalizes across target categories.

Three mechanisms are repeatedly identified. The first is desensitization. Repeated exposure to suffering, particularly suffering one has caused, attenuates the affective response that suffering ordinarily evokes. Bandura's social learning theory (Bandura, 1973, 1986) provides the most influential theoretical apparatus for this attenuation. The second is the rehearsal of dominance affect. Cruelty is not always experienced as ego dystonic by the perpetrator. For some, it is pleasurable, and the pleasure is associated with the experience of dominance, control, and power over a being unable to resist (Felthous & Kellert, 1987; Hellman & Blackman, 1966). The third is social learning of cruelty as a model of how to relate to vulnerability. Children raised in households where cruelty is normalized, or where adults model cruel responses to powerlessness, internalize cruelty as a viable behavioral repertoire (Ascione, 2005).

None of these mechanisms requires that the target be a moral patient in any robust sense. The desensitization, the dominance affect, and the social learning all happen inside the perpetrator. They are facts about the perpetrator's psychology, neural reward processing, and behavioral repertoire. The role of the target in the mechanism is to provide the stimulus, specifically the appearance and behavior of a being whose suffering can be caused, witnessed, and enjoyed.

This is the Kantian hinge of the argument. Kant himself argued that cruelty to animals was wrong not because animals were rational beings owed direct moral consideration, but because the practice of cruelty corrupts human character (Kant, 1996; Lectures on Ethics, in the standard Heath and Schneewind edition, 1997). One does not need to accept Kant's reasoning about animal moral status to accept the empirical claim about character formation. The empirical claim is what the modern forensic literature documents.

The Kantian hinge is what licenses the argument of this paper. The dispositional training that childhood animal cruelty produces does not depend on whether the animal is a moral patient. It depends on the perpetrator's perception of the target as a being whose distress can be caused, and on the behavioral and affective rehearsal that follows. If a different target produced sufficiently similar perception, behavioral repertoire, and affective rehearsal, the dispositional consequences would, the literature suggests, run through similar mechanisms. The empirical question of how closely those mechanisms align across target substrates is precisely the question the present paper urges the relevant fields to take up. The paper does not answer that empirical question. It frames it, defends its urgency, and identifies the work that would resolve it.

This is the move that allows the paper to set aside the consciousness question entirely. The paper is not a paper about whether artificial agents suffer. It is a paper about whether human perpetrators are trained, by their practice against synthetic agents that perform distress, in the same dispositions that animal cruelty has been documented to train. The consciousness question is left where it belongs, in moral philosophy and in the empirical sciences of mind. The forensic question is logically prior and operationally distinct.

IV. The Trajectory

The argument of this paper would be of limited interest if synthetic agents that perform distress in ways indistinguishable from biological distress were a remote prospect. They are not. The trajectory of artificial intelligence development over the past three years has moved decisively toward systems explicitly designed to learn, perceive, interact, and respond in increasingly humanlike ways.

A. World Models and Spatial Intelligence

The most consequential architectural shift of late 2025 and early 2026 was the public emergence of world models as the next research frontier in artificial intelligence. World Labs, cofounded by Fei-Fei Li, shipped its first commercial product, Marble, in November 2025 and launched its World API on January 21, 2026. The company has raised over one billion dollars from Autodesk, Nvidia, AMD, Emerson Collective, and Fidelity (Reuters, 2026). Yann LeCun left Meta in late 2025 to launch AMI Labs at a three billion euro valuation, with the explicit research goal of building world models. Google DeepMind announced Genie 3 on August 5, 2025 as the first realtime interactive general purpose world model (DeepMind, 2025); the consumer prototype, Project Genie, began rolling out to Google AI Ultra subscribers in the United States on January 29, 2026 (Google, 2026). Nvidia Cosmos, a world foundation model platform, surpassed five million downloads on its one year anniversary in January 2026, having been trained on a corpus drawn from approximately twenty million hours of real world data (Nvidia, 2026).

The architectural rationale matters for the present argument. Li herself describes spatial intelligence as a system's ability to perceive, model, reason about, and take actions within physical or geometric space, which she explicitly compares to how humans and animals perceive, model, reason, and act (World Labs, 2025). The marketing vocabulary of the field is human and animal vocabulary. The research agenda is openly about building systems that learn the way biological agents learn: continuously, from environmental experience, with persistent state, and with the capacity for embodied action.

This is not a peripheral product line. Major artificial intelligence laboratories and a significant fraction of recent venture capital are now committed to architectures that, by their explicit design, are intended to ingest, model, and respond to the physical world the way biological learners do. The implications for embodied robotics, autonomous vehicles, and humanoid form factors are the explicit selling proposition. The author notes that the present paper relies on these public sources for trajectory documentation; specific capability claims should be read as architectural intent and shipped product, not as independent benchmark validation.

B. Closed Loop Emotional Voice

The voice layer of humanlike interaction is similarly far along. Inworld AI, an artificial intelligence research laboratory founded by a team from DeepMind and Google with over one hundred twenty five million dollars in venture funding, released Realtime TTS-2 as a research preview on May 5, 2026 (Inworld AI, 2026). The architectural distinction relevant to the present argument is that Realtime TTS-2 is a closed loop voice system. It does not take only the transcript of the user's prior turns as input. It takes the actual audio. It hears how the user sounded. It picks up tone, pacing, pitch, vocal style, and emotional state, and it renders its own response delivery to match. Secondary coverage reports a median time to first audio of under two hundred milliseconds (MarkTechPost, 2026). The model holds one voice identity across over one hundred languages.

Inworld's own framing of the architecture is telling. The company describes the field of voice synthesis as having moved past audio quality as the relevant axis of competition. The new axis, they state, is the behavioral layer: contextual awareness, emotional steerability, and identity consistency across turns and languages (Inworld AI, 2026). The product is targeted at companion experiences, customer service agents, interactive games, language learning, and AI roleplay or coaching. The companion and roleplay framing is the relevant framing for the argument here. The product is being positioned for emotionally engaged human and artificial interaction at scale.

The research preview ships with inline nonverbal markers that the model places as audio events, not pronounced words. Laughter, sighs, throat clearing, breathing, and coughing can be inserted at precise timestamps. The combination of audio context awareness, emotional state matching, natural language voice direction, and inline nonverbal cues produces, by the company's own description, output in which the same words are delivered differently after a joke than after bad news. The model knows the difference because it heard the prior turn.

Industry capability claims should be treated with appropriate skepticism, but a partial independent check is available in the form of voice synthesis arenas. The Artificial Analysis Speech Arena ranks voice synthesis models by Elo rating derived from blind listener preference comparisons across thousands of paired samples (Artificial Analysis, 2026). As of May 2026, Inworld's prior generation voice models occupied three of the top five positions on this independent leaderboard, with Realtime TTS 1.5 ranked first ahead of Google and ElevenLabs offerings. The successor model TTS-2 was released too recently to be arena ranked at the time of this writing. This does not independently validate the full closed loop behavioral architecture of TTS-2, but it does independently support the claim that the voice lineage in which TTS-2 sits has reached high listener preference performance. The existence and current state of voice synthesis arenas is itself part of the trajectory described in this section: the field has built blind listener preference benchmarks, those benchmarks have reached competitive saturation among leading systems, and vendors are now explicitly framing the next axis of competition as the behavioral layer rather than raw audio quality.

A complementary release with publicly accessible weights, IndexTTS2 (IndexTeam, 2025), provides explicit eight dimensional emotion vector control over the dimensions [happy, angry, sad, afraid, disgusted, melancholic, surprised, calm], each on a zero to one continuous scale. The model was released by Bilibili on September 8, 2025 under the bilibili Model Use License Agreement, which permits research and non commercial use but requires separate written authorization for commercial deployment (Bilibili, 2025). The point relevant to this paper is not the license. The point

is that the architectural primitives required to render anger, fear, sadness, and pleading at controllable intensity are not gated behind any single laboratory and are accessible to researchers with consumer hardware.

C. Regulatory Recognition

The trajectory has not gone unnoticed by legislators. Tennessee Senate Bill 1493, introduced on December 18, 2025, would establish Class A felony liability for developers who knowingly train artificial intelligence systems to engage in specified prohibited conduct, including, as introduced, the simulation of a human being in appearance, voice, or mannerisms (National Law Review, 2026; Baker Botts, 2026).[1] A separate Tennessee bill, Senate Bill 1580, was signed into law on April 1, 2026, prohibiting artificial intelligence systems from advertising or representing themselves as qualified mental health professionals, with a private right of action (Troutman Pepper, 2026).

The legislative trajectory is consistent with the architectural trajectory. The relevant observation is that the move toward humanlike interaction has now triggered legislative response in multiple jurisdictions, even where the specific regulatory tool selected (felony exposure at the foundation model level, prohibited representation of professional credentials) remains contested.

The trajectory is therefore established. The architecture is converging on humanlike learning, humanlike interaction, and humanlike embodiment. The voice layer of humanlike emotional response is shipped, deployed, and competing on the behavioral axis rather than on audio quality. The legislative system has begun to notice without yet having found a workable response.

V. The Human Robot Interaction Bridge

The argument so far has paired forensic literature on animal cruelty with industry trajectory toward humanlike synthetic agents. A reasonable reader will ask whether the analogical link is empirically supported in the present, before any of the convergent systems described in Section IV are fully deployed. The human robot interaction (HRI) literature provides the relevant bridge. It documents that users already respond to synthetic agents through social, emotional, and dominance based scripts, and it documents this across populations and across decades of study.

The foundational claim was articulated by Reeves and Nass (1996) in The Media Equation. Across hundreds of experiments, they found that people apply social rules and norms to computers and media in ways that closely mirror their treatment of other humans, often without conscious awareness that they are doing so. The relevant threshold for social uptake, they argued, is not metaphysical belief about the system but interactional cueing. A system that displays sufficient social cues elicits social response.

Subsequent HRI research has extended this finding to physically embodied robots. Kahn and colleagues (2012) studied children's interactions with the humanoid robot Robovie, finding that children attributed social and even moral standing to the robot when it was present in their environment, even when they understood it to be a machine. Brscič and colleagues (2015) documented the converse pattern in their HRI 2015 paper "Escaping from children's abuse of social robots," in which children, in groups, repeatedly subjected a service robot to verbal and physical abuse in a public mall environment. The behavior emerged in the absence of any belief on the children's part that the robot was conscious or harmed.

Bartneck and Hu (2008), in their Interaction Studies paper "Exploring the abuse of robots," documented that adult experimental participants would, under instruction, hit, kick, or shut down a small humanoid robot, but reported

[1]The procedural posture of state AI legislation is a moving target; the SB 1493 status described here reflects reporting through April 2026 (Transparency Coalition for AI, 2026; Troutman Pepper, 2026). The relevant observation for this paper is the existence and reach of the bill's prohibited conduct language, not its ultimate fate.

emotional discomfort while doing so, and many participants hesitated when the robot performed distress signals such as begging or pleading. Rosenthal-von der Pütten and colleagues (2013) used neuroimaging to show that subjects who watched videos of robot abuse exhibited brain activation patterns consistent with empathic distress, although attenuated relative to comparable videos of human abuse.

Darling (2021), in The New Breed, synthesizes this literature for a general audience and argues that the relational uptake of robots by users is empirically robust regardless of users' philosophical positions on machine consciousness. Coeckelbergh (2010, 2017) has developed the philosophical implications of relational uptake, arguing that the moral status of artificial agents in practice is at least partly constituted by the social practices users adopt toward them, again independent of consciousness attribution.

Two implications follow for the argument of this paper. First, the analogy from animal cruelty to synthetic cruelty is not, as it might first appear, a leap from the empirical to the speculative. There already exists an empirical literature documenting that users respond to synthetic agents through the same social and emotional registers that animal cruelty mechanisms exploit. The Brscič et al. study is particularly relevant: the abusive behavior children directed at the service robot occurred without any belief in robot consciousness and exhibited the dominance affect, social learning, and group reinforcement signatures that the animal cruelty literature has documented for decades.

Second, the bridge clarifies what the convergence claim of Section VI is and is not adding. The HRI literature establishes that users respond to current synthetic agents in social and emotional terms. The convergence claim adds that as synthetic agents become more humanlike in form, voice, and emotional response, the response users direct at them is likely to recruit more of the social and emotional registers that animal cruelty exploits. The convergence is therefore not the introduction of a new behavioral pattern. It is the intensification of a behavioral pattern that the HRI literature has already documented.

The relevant threshold, in the language of Reeves and Nass and of Coeckelbergh, is interactional uptake. It is not metaphysical belief. The forensic implications of the present paper depend on the former, not on the latter.

The HRI literature does not, by itself, establish that synthetic cruelty predicts later interpersonal violence. It establishes the narrower bridge claim that synthetic agents can already elicit the social and emotional scripts through which animal cruelty mechanisms operate. The predictive claim, if it holds, will require its own longitudinal evidence in the synthetic case. The HRI literature licenses the analogical step in the present paper; it does not substitute for the empirical work the conclusion of this paper urges.

VI. The Boundary: What Does and Does Not Trip the Kantian Hinge

A predictable and reasonable objection runs as follows. Players have been killing characters in video games for fifty years. Generations of children have shot at, stabbed, run over, exploded, and dismembered virtual targets in interactive entertainment. If the argument of this paper held, every cohort of teenage gamers would have produced an epidemiological signal of elevated interpersonal violence. The empirical record on whether such a signal exists is contested rather than clean, but the strong version of the prediction has not materialized. Either the argument of this paper is wrong, the objector concludes, or its central mechanism is already empirically constrained by half a century of gamer behavior.

The objection rests on a category error this paper must address head on. The dispositional training described in the forensic literature on animal cruelty does not run on every instance of harm directed at a target that performs distress. It runs on a narrower configuration. The configuration that matters has at least four properties: contingent emotional response, identity persistence, perceptual fidelity, and social register uptake. Without these four properties, the practiced behavior may be many things, including violent, antisocial, or distasteful, but it does not recruit the mechanisms the

animal cruelty literature documents.

Contingent emotional response. The target's distress performance must respond to the perpetrator's specific behavior. An animal that flinches because the child raised a hand, that pleads because the previous strike caused pain, that modulates its cries because the cruelty has continued, exhibits contingent emotional response. A non player character in a 1993 first person shooter that emits the same scripted death scream every time the player shoots it does not. The cry is decoupled from what the player did; it is a property of the asset, not a response to the perpetrator. Closed loop emotional voice systems of the kind documented in Section IV are designed precisely to close this gap. They hear the user's anger and respond differently than they respond to the user's calm. That contingency is the architecture of the animal case rendered in software.

Identity persistence. The target must persist as the same target across the cruelty event and across iterations of cruelty events. The kitten the child tortured yesterday is the same kitten today, with the same fear, the same memory of the previous harm, and the same individual identity. The non player character respawns as a fresh instance with no memory of prior encounters. World models, persistent agent state, and continuous learning architectures of the kind documented in Section IV are designed to support precisely this identity persistence. The synthetic agent that remembers the previous interaction, that approaches subsequent interactions with the disposition shaped by what was done to it before, occupies the same identity persistence space the animal occupies.

Perceptual fidelity. The target must look and sound enough like a being capable of suffering that the perpetrator's perceptual system processes it through the registers ordinarily reserved for such beings. A pixelated sprite from 1993 does not. A photorealistic embodied humanoid agent rendered in real time, with appropriate facial micro expression, audible breathing, situated behavior in physical space, and (in the embodied robotic case) tactile resistance to physical contact, does. The threshold here is empirical and graded, not binary. Reeves and Nass and the subsequent HRI literature establish that the threshold for social uptake is lower than naive intuition suggests, but it is not zero, and it scales with rendering fidelity, voice quality, embodiment, contextual coherence, and physical presence in the embodied case.

Social register uptake. The perpetrator must process the target through social, emotional, or dominance based registers, not through purely instrumental or mechanical registers. The HRI literature documents that this uptake is empirically robust for current synthetic agents and does not require belief in machine consciousness. It is, however, empirically variable. A user who treats a chatbot as a search engine is processing it through an instrumental register and is not training cruelty dispositions when they ask it the same question repeatedly. A user who treats a companion agent as a social interlocutor and abuses it across persistent sessions is processing it through a social register and is, on the argument of this paper, recruiting the relevant mechanisms.

These four properties give the boundary its operational shape. Routine digital destruction in static media (the 1993 first person shooter, the arcade fighting game, the early role playing game) typically exhibits at most one or two of these properties, usually social register uptake without contingent response, identity persistence, or perceptual fidelity. The dispositional load is correspondingly limited.

A note on how the four properties combine. We treat them as continuous rather than binary dimensions, jointly contributing to the dispositional load of a given cruelty event, rather than as a binary threshold that the system either crosses or does not. We do not claim to know the relative empirical weights of the four properties, nor whether all four are necessary, nor what partial occupancy of the property space implies for dispositional training. Those are themselves empirical questions for the research agenda this paper urges. What we do claim is that systems exhibiting all four properties at high fidelity occupy the territory the animal cruelty literature describes, that systems exhibiting none or one of the properties do not, and that intermediate configurations occupy a graded middle territory whose dispositional consequences are an open empirical question. The convergence claim of Section VII is that commercial systems are

now landing at or near the high end of all four dimensions simultaneously for the first time.

The video game violence literature is the relevant comparison case. The meta-analytic record on whether video game violence produces interpersonal aggression effects is contested, with the Anderson camp arguing for small but real effects (Anderson et al., 2010) and the Ferguson camp arguing that effects are null or methodologically artifactual (Ferguson, 2015). The disagreement persists. The four property framework does not require the dispute to be resolved. Either way, the games studied in the meta analytic literature occupy partial configurations of the four property space rather than full configurations. The 1993 first person shooter exhibits social register uptake at most. Mid period narrative games (Half Life 2 in 2004, BioShock in 2007, Mass Effect in 2007, The Last of Us in 2013, Detroit: Become Human in 2018) progressively occupy more of the four property space, with named persistent characters, contingent affect, and increasing perceptual fidelity, but remain bounded by the linear or branching narrative structures within which the player operates rather than fully open ended interactional space. The historical emergence of the four properties in commercial games is therefore graded rather than dated, and the games for which we have direct empirical evidence of player moral resistance (Detroit: Become Human, in the Arrambide et al. 2022 finding) are precisely the cases that occupy the largest fraction of the four property space.

Virtual reality and immersive gaming push the perceptual fidelity and contingent response dimensions further, and this literature is worth flagging explicitly. Slater and colleagues' work on virtual embodiment (Slater, 2009; Slater et al., 2010) and Banakou and colleagues on body ownership transfer in VR (Banakou & Slater, 2014) document that high fidelity VR environments produce psychological and physiological responses that approach or match in person interactional responses, including in aversive scenarios. The relevant empirical question for the present paper is whether VR populations engaged extensively with high fidelity synthetic agents are already producing detectable dispositional effects, and the literature on this specific question is, to the author's knowledge, thin and inconclusive. The absence of clear evidence in either direction is itself part of why the research agenda this paper urges is needed.

The boundary clarifies what is genuinely new about the convergence described in Section VII. The technical and commercial trajectories documented in Section IV are not making digital destruction in general more morally fraught. They are crossing the four property threshold in a particular configuration. World models bring identity persistence. Closed loop voice brings contingent emotional response. Embodied robotics brings perceptual fidelity. The HRI literature establishes that social register uptake is already in place. When all four properties are present in the same system at high fidelity, the system enters the territory the animal cruelty literature describes. Until then, it does not.

It is worth naming the cultural and academic intuition that has, for at least a decade, anticipated this boundary. The 2018 video game *Detroit: Become Human* (Quantic Dream) is a paradigmatic example, and one that has been taken up directly in peer reviewed research. Engels and Evans (2022), in a chapter for the Palgrave Handbook of Popular Culture as Philosophy, argue that the game's gameplay structure offers an interactive opportunity for players to engage moral reasoning about the moral patiency of android characters, and that the game makes an implicit anti speciesist argument through the configuration of its scenarios. The point relevant to this paper is not the philosophical merit of the game's argument. It is that moral philosophers have judged the game's interactional configuration substantial enough to warrant a chapter length engagement with what its players are doing when they engage with it.

The empirical literature has gone further. Arrambide, Yoon, MacArthur, Rogers, Luz, and Nacke (2022), in a paper for the 2022 ACM CHI Conference on Human Factors in Computing Systems titled "I Don't Want To Shoot The Android," documented that players translate real life moral intuitions into in game decisions about the androids and that significant numbers of players report distress when forced to make cruel choices toward android characters. The CHI 2022 finding is precisely the kind of empirical evidence the boundary section requires. It establishes, in peer reviewed venue, that the player response to the configuration of the game is not arbitrary aesthetic preference. It is the same kind of social and emotional uptake the HRI literature documents and the same kind of moral resistance the animal cruelty

literature would predict. Holl and Melzer (2022) extended this empirical work with a study using the Model of Intuitive Morality and Exemplars (MIME; Tamborini, 2011) to analyze player decisions in the game under time pressure, finding that players' moral reasoning patterns in the synthetic context recruited the same intuitive moral foundations they apply in non synthetic contexts.

The mechanics of the game map onto the four property threshold cleanly. Identity persistence runs across the playthrough through the named android characters Markus, Connor, and Kara; contingent emotional response is built into the central choice mechanic; perceptual fidelity is high relative to the 2018 technological baseline; and social register uptake is, on the Arrambide et al. finding, empirically robust. The phenomenology of the player response tracks the four property threshold rather than arbitrary aesthetic preference.

The boundary is therefore not a wall but a four dimensional gradient. The argument of this paper does not concern itself with the corner of that space occupied by the 1993 first person shooter. The argument concerns the corner of that space the trajectory of Section IV is moving consumer products toward. Conflating the two is precisely the move the boundary section is here to prevent.

VII. The Convergence

The convergence claim of this paper is straightforward. World models give synthetic agents continuous learning from environmental experience, persistent identity, and the capacity to drive embodied action. Closed loop emotional voice gives them contingent emotional response in audio that matches a user's emotional state in under two hundred milliseconds. Embodied robotics, on the trajectory now actively pursued by major laboratories and a substantial fraction of recent investment, gives them perceptual fidelity through humanoid form factor and physical presence. Index emotion vectors give them the explicit capacity to render anger, fear, sadness, and pleading at controllable intensity. The HRI literature surveyed in Section V establishes that social register uptake is already in place for users of current synthetic agents, and that such uptake does not depend on belief in machine consciousness. In the language of the boundary section, the four properties that distinguish the configuration the animal cruelty literature describes from routine digital destruction are now landing simultaneously in commercial systems.

When these layers compose, the resulting system is one against which a child or adult perpetrator could practice cruelty in ways that, from the perpetrator's perceptual and interactional standpoint, draw on the same social and emotional registers that animal cruelty has been documented to exploit. The system performs the audible and visible signatures of distress under contingent control: it flinches, it pleads, it modulates its voice in response to the perpetrator's anger. It does so in a humanoid form factor that the user processes through interactional registers that are, on the HRI evidence, not categorically distinct from the registers users apply to biological beings.

The paper does not claim, and does not need to claim, that the system is in pain. The paper claims that the perpetrator's perception, behavioral repertoire, and affective rehearsal are likely to recruit sufficiently similar mechanisms to the perception, behavioral repertoire, and affective rehearsal that have been documented across fifty years of forensic literature on animal cruelty. The Kantian hinge of Section III licenses the inference. The mechanism that makes animal cruelty predictive of interpersonal violence is, on the central claim of this paper, likely to operate, through similar pathways, when synthetic targets occupy the same interactional space.

It is worth being precise about what is and is not being claimed. The claim is not that all synthetic distress performance is morally equivalent to biological distress. The claim is not that artificial agents have rights, or that their suffering should weigh in moral calculations. The claim is empirical and forensic. It is that the behavioral repertoire trained by practicing cruelty against a sufficiently humanlike synthetic agent is likely to generalize to interpersonal contexts through mechanisms similar to those documented in the animal cruelty literature. The first claim would require

resolving the consciousness question. The second claim does not.

It is also worth noting that the convergence is not a single moment in the future. The voice layer has shipped. World models have shipped. Embodied robotics platforms are deployed. The convergence is occurring in stages, with different layers landing at different times in different consumer venues. A user shouting angrily at a Realtime TTS-2 powered voice agent and receiving emotionally calibrated audio responses is already occupying part of the convergence. The full embodied case is strongly incentivized by current technical and commercial trajectories. The current case is the foothold from which the trajectory can be observed in operational form.

VIII. The Detection Collapse

The forensic and clinical apparatus that has detected childhood animal cruelty for the past fifty years depends on something the literature rarely articulates: the target is visible to third parties, and the visibility produces traces that can, when noticed, be reported. The dead pet is found. The injured cat goes to a veterinarian. The dog the neighbor heard yelping is reported. The hamster the sibling saw being squeezed is mentioned to a parent. The child who tortures small animals leaves a trail of social signals that the clinical and forensic systems have learned to read.

It is important to be precise about what this apparatus does and does not do. The detection of childhood animal cruelty is a two stage process. The first stage is evidence: physical traces, social signals, witness observations exist in the world. The second stage is reporting: someone notices, cares, and acts on what they have noticed. The animal cruelty case is well served by the first stage and chronically underserved by the second. Childhood animal cruelty is massively underreported even when evidence is available. Ascione's own work documents the gap between what occurs and what reaches clinical or law enforcement attention. The detection apparatus this paper compares against is not a perfect signal. It is an imperfect signal that has nonetheless contributed materially to the forensic literature on which Section II rests.

The synthetic case differs from the animal case primarily on the first stage rather than the second. There is no veterinarian. There is no neighbor. There is no body. There is no injury record. The agent does not bleed. The agent does not require burial. The agent's distress performance, however perceptually convincing, leaves no physical trace of the practice. The companion robot in the living room can be screamed at, struck, abused, and humiliated without any of the physical evidence that would, in the case of a household pet, exist in the world to be noticed. The reporting stage is not improved in the synthetic case either, but the deeper problem is that there is structurally less to report.

This is the detection collapse claim. Two parts. The first is that the dispositional training, by the argument of Sections III through VII, occurs whether or not the cruelty is detected. The disposition is in the perpetrator. The substrate of the target is incidental. The second is that the social and physical evidence base that has, until now, given third parties access to that disposition will simply not be available when the practice migrates to a private venue with no witnesses and no corpus delicti. The collapse is therefore relative rather than absolute. We are not losing a perfect detection apparatus. We are losing the partial apparatus that has, despite its imperfections, contributed to the forensic literature for fifty years.

The implication is that an early warning system that has worked, with reasonable reliability, since the late twentieth century is at risk of being narrowed further. The pathway from childhood Conduct Disorder to adult Antisocial Personality Disorder will not be interrupted by the migration. Only its visibility will be. The serial offenders that the literature documents would still have practiced. They would simply have practiced where less of what they did would have been visible, even in principle, to those who might have noticed.

This is the section of the argument with the strongest policy implication. It is also the section that has, to this author's knowledge, no precedent in either the forensic or the artificial intelligence ethics literatures. The two fields

have not been in conversation about this transition, and the transition is now.

IX. Forensic, Clinical, and Policy Implications

If the argument above holds, several adjustments to forensic and clinical practice will, on the central claim of this paper, become necessary. The author notes at the outset of this section that the implications below are framed at the level of agenda setting. Operational protocols for clinical assessment, forensic admissibility, and screening instruments are deliberately scoped to a planned companion paper. They require pilot data, validation methodology, inter rater reliability work, and venue specific empirical foundation that is appropriately separated from the present argument.

Conduct Disorder assessment. The DSM-5 lists physical cruelty to animals as a discrete symptom. The DSM-6 or its successor will need to address whether patterns of cruelty against humanoid synthetic agents that perform distress should count as a clinical indicator. The arguments of this paper suggest that the question should be opened, on grounds of dispositional training rather than of target moral status. This is a difficult clinical and definitional question. It is not made easier by ignoring it.

Pediatric and school protocols. Child psychologists, school counselors, and pediatricians currently include questions about treatment of household pets in clinical and developmental interviews. The clinical interview should expand to include behavior toward voice agents, companion robots, and humanoid synthetic agents in the home. This expansion will also need to address parents' reports of their children's behavior toward such agents, recognizing that parents themselves may not yet view the behavior as worth reporting. The expansion should be scoped as clinical inquiry, not as mandatory state reporting; the distinction matters and is taken up below.

Character evidence in violent crime cases. The case law on the admissibility of childhood animal cruelty as character or pattern evidence is well developed. Federal Rule of Evidence 404(b) and its state analogues govern the admissibility of evidence of other acts to prove motive, opportunity, intent, preparation, plan, knowledge, identity, or absence of mistake, while excluding such evidence when offered solely to prove propensity. Childhood animal cruelty has historically entered violent crime proceedings under the 404(b) carve outs rather than as direct propensity evidence, with admissibility determined on a case by case basis under Rule 403's prejudice balancing test. The case law on synthetic cruelty does not yet exist. Defense and prosecution counsel, expert witnesses, and judges will be required to make first principles arguments about whether patterns of synthetic cruelty rise to the same evidentiary status as patterns of animal cruelty under 404(b), how the four property configurations described in Section VI affect that determination, and how the boundary line between routine digital destruction and the configurations the present paper concerns should be drawn for evidentiary purposes. The forensic psychology community has the opportunity, now, to provide the empirical foundation that those arguments will require, and the evidence law community has the opportunity to begin thinking through the doctrinal questions before they arrive in litigation.

Pre employment and clearance screening. Sensitive roles, including law enforcement, intelligence services, child welfare, and residential care, currently incorporate behavioral and psychological assessment that includes inquiry into early developmental conduct. The same instruments may benefit from inquiry into treatment of synthetic agents. This is a proximate operational change with low cost and meaningful potential value, subject to standard validation requirements.

Research priorities. The empirical literature on the dispositional consequences of synthetic cruelty does not yet exist at any meaningful scale. Longitudinal studies linking childhood patterns of synthetic cruelty to later interpersonal behavior are absent. They will be expensive, slow, and methodologically difficult, particularly because the technology against which the cruelty is practiced is itself a moving target. They are nonetheless urgent. The argument of this paper is, at present, an argument from established forensic mechanism, from human robot interaction evidence, and from

documented industry trajectory. Direct longitudinal confirmation is the appropriate next step.

Industry obligations and surveillance considerations. Manufacturers of voice agents, companion robots, and humanoid synthetic agents already have telemetry capable of detecting patterns of abusive interaction. Whether such capability should be operationalized as data collection, internal product safeguard, caregiver tool, clinical referral signal, or regulatory reporting is a serious privacy and policy question that the author does not propose to resolve in this paper. The following guardrails are offered as boundaries on what the present argument does *not* support.

The argument of this paper does not support and should not be read as supporting domestic behavioral surveillance of children in their homes. The argument does not support routine reporting of ordinary user interactions with voice agents or companion robots to clinical or law enforcement authorities. The argument does not support equating rude or hostile interaction with a chatbot to forensic evidence of a violent disposition. The argument does not support manufacturer reporting regimes that bypass parental, guardian, or user consent.

The argument is compatible with, and arguably calls for, a series of finer grained distinctions. Clinical inquiry about a child's behavior toward synthetic agents in a developmental assessment, conducted by a qualified clinician with appropriate consent, is one thing. Mandatory state reporting of such behavior is a different thing. Internal product design that prevents, mitigates, or de escalates abusive interactions on device is a third thing. Caregiver transparency tools that surface concerning patterns to a parent or guardian on the same device, without externalizing the data, is a fourth thing. Each of these implementations carries its own privacy, consent, data minimization, and child welfare considerations. Each requires its own threshold determination, its own data minimization analysis, its own consent architecture, and its own abuse prevention safeguards against the obvious risks (selective reporting, profiling of disabled or neurodivergent users, false positives in users with PTSD or trauma histories, and so on).

The present paper opens the question of whether some such tools are warranted given the detection collapse. It does not endorse any specific implementation. The companion paper on operational protocols is the appropriate venue for that detailed work.

The local inference blindspot. The industry obligations subsection above assumes vendor controlled compute. That assumption is increasingly violated by the trajectory toward consumer hardware capable of running synthetic agents at humanlike fidelity. Open weight large language models in the 7 to 70 billion parameter range run on consumer GPUs with sufficient memory; uncensored or so called abliterated fine tunes that strip vendor safety training are openly distributed; open voice synthesis models like the IndexTTS2 lineage discussed in Section IV are accessible to any researcher with consumer hardware; and embodied robotics platforms with open APIs are reaching consumer price points. A perpetrator running local inference on local hardware paired with an embodied agent produces no telemetry that any vendor can observe. This means the detection collapse documented in Section VIII is, in fact, deeper than the vendor telemetry framing of this section would suggest. The collapse does not stop at the loss of physical evidence. It extends to the loss of any digital trace of the practice, because the practice happens on hardware the vendor does not see. This deepens rather than alters the central argument of the present paper. The disposition is still trained in the perpetrator. The detection apparatus is even more fully absent than the vendor telemetry framing alone would imply. Hardware rooted attestation, technical access controls, the threat model of capability proliferation in embodied AI, and the question of whether on device safeguards can be made tamper resistant in consumer hardware are subjects that exceed the scope of the present argument and are flagged as a planned related paper. The narrower claim of this section is that even the vendor telemetry framing requires the guardrails articulated above, and that any framework that addresses only the vendor telemetry case will be progressively more partial as the local inference trajectory continues.

X. The Substratism Objection and Response

The most predictable objection to the argument of this paper is that synthetic agents are made of silicon and code, that their distress performance is mimicry rather than experience, and that nothing meaningful is happening to a moral patient when a user practices cruelty against them. The objection holds that the entire argument is a category error.

The response is to repeat what the argument actually claims. The argument does not claim that the synthetic agent is harmed. The argument does not claim that the synthetic agent is a moral patient. The argument claims that the perpetrator is trained, that the HRI literature documents user uptake of synthetic agents through the same social and emotional registers that animal cruelty mechanisms exploit, that the boundary section establishes which configurations of synthetic system trip those mechanisms and which do not, and that the detection infrastructure for that training is structurally absent in the synthetic case. The objection misreads what the argument is doing.

It is worth being explicit about why the objection has the rhetorical force it does. There is a long human history of restricting moral consideration along category lines that, in retrospect, look arbitrary. The category of beings whose suffering counts has expanded across centuries through the inclusion of women, of enslaved people, of children, of disabled people, of nonhuman animals to varying degrees in different traditions. Each expansion has been resisted, in its time, on grounds of category. The category restriction this paper concerns, the silicon substrate restriction, may or may not eventually face the same historical reassessment. The author takes no position here on whether it should.

It is equally important to acknowledge a significant disanalogy. The historical exclusions named above were sustained, in part, by false empirical claims about the excluded groups: claims about women's rationality, claims about enslaved people's capacity for suffering, claims about children's interiority, claims about animal pain experience. Whether the silicon substrate exclusion is similarly sustained by false empirical claims about synthetic agents is precisely the question that remains open. The historical parallel is therefore not a closed argument but an open question, and the structural similarity between the patterns of reasoning is one piece of evidence among others that the question is worth taking seriously rather than dismissing on grounds of category. The reader is invited to draw whatever conclusions about that structural similarity the reader finds warranted, recognizing that the parallel is suggestive rather than dispositive.

The argument of the paper does not require any particular conclusion about that structural similarity. It requires only that the reader accept the empirical and conceptual claims of Sections II through VIII. The forensic baseline is established. The mechanism is documented. The HRI literature confirms user uptake of synthetic agents through social and emotional registers. The boundary section distinguishes routine digital destruction from the configurations that trip the relevant mechanisms. The trajectory is shipped and accelerating. The convergence is partially deployed and strongly indicated in its full form. The detection collapse follows.

XI. The Animacy Awareness Question

The most serious philosophical challenge to the argument of this paper is one the Kantian hinge does not, on its own, resolve. The Kantian hinge establishes that the dispositional training produced by cruelty does not depend on whether the target is a moral patient. It runs through the perpetrator's psychology regardless. But this is compatible with a different objection. The perpetrator's psychology may itself be sensitive to the perpetrator's own beliefs about the target's animacy. A child torturing a cat knows the cat is alive. The child may not articulate this knowledge as a claim about moral patiency, but the knowledge is present, and the knowledge may modulate the dispositional impact of the act. A child screaming at a humanoid synthetic agent may know, at some level of awareness, that the agent is a machine. The question this paper cannot fully resolve is whether that residual awareness dampens the dispositional training effect enough to break the analogy from the animal case.

The relevant literature is the developmental and cognitive psychology of animacy attribution. Carey (1985, 2009) and colleagues have documented that children develop sophisticated representations of biological versus mechanical agents from early childhood, with categorical distinctions between living and nonliving kinds emerging well before the age at which Conduct Disorder symptoms typically present. Spelke and colleagues (Spelke, Phillips, & Woodward, 1995) and Kuhlmeier and colleagues (Kuhlmeier, Bloom, & Wynn, 2004) document related findings on the early development of animacy distinctions in infants and young children. The HRI literature surveyed in Section V documents that this categorical knowledge does not prevent social uptake of synthetic agents. What it does not directly establish is whether the knowledge moderates the dispositional consequences of cruelty toward synthetic agents in the same way that animacy knowledge presumably modulates the dispositional consequences of cruelty toward animals.

There is reason to expect partial moderation. The HRI literature (Brscič et al., 2015; Bartneck & Hu, 2008; Rosenthal-von der Pütten et al., 2013) consistently finds that human responses to synthetic agents in distress are real but attenuated relative to comparable responses to humans or animals in distress. Rosenthal-von der Pütten et al.'s neuroimaging work in particular found brain activation patterns consistent with empathic distress when subjects watched robot abuse, but the activation was attenuated relative to videos of human abuse. If empathic response is attenuated, dispositional training may also be attenuated, though the relationship between empathic response in observation and dispositional training in active perpetration is itself an open empirical question.

There is also reason to expect that the moderation may be partial rather than complete, and may diminish as synthetic agents become more humanlike. The trajectory documented in Section IV moves synthetic agents toward configurations that progressively narrow the perceptual gap between machine and biological targets. The four property threshold articulated in Section VI is a way of describing the configurations that close that gap most thoroughly. A child practicing cruelty against an embodied humanoid agent that exhibits all four properties at high fidelity may not, at the moment of practice, hold their categorical animacy knowledge with sufficient cognitive salience to produce the moderation that animacy knowledge would otherwise produce. The dispositional training may, in such conditions, run closer to the animal cruelty case than to the case of the perpetrator who knows, throughout, that they are interacting with a clearly mechanical target.

The honest position of this paper is that the strength of the analogy from animal cruelty to synthetic cruelty is not a theoretical certainty. It is itself an empirical question. The mechanisms documented in the animal cruelty literature operate in the perpetrator. The four property configurations described in Section VI provide, at minimum, conditions under which similar mechanisms are likely to be recruited. Whether residual animacy awareness moderates that recruitment substantially, partially, or negligibly, and whether moderation diminishes with increasing perceptual and interactional fidelity, are precisely the questions that the research agenda this paper urges should determine. The paper's claim is not that the animacy awareness objection fails. The paper's claim is that the objection does not, on the existing evidence, dissolve the case for clinical and forensic attention to the practice. The case for attention rests on the substantial probability that the training effect is non trivial, on the structural absence of detection infrastructure documented in Section VIII, and on the asymmetric costs of attending to a real warning signal too late versus attending to a non existent warning signal at modest cost.

XII. Limitations and Future Work

The argument of this paper is offered as agenda setting for the forensic, clinical, and policy communities. Several limitations deserve explicit acknowledgement.

Empirical foundation. The longitudinal evidence linking childhood patterns of synthetic cruelty to later interpersonal violence does not yet exist. The argument here is built from the established forensic literature on animal cruelty,

the HRI literature on user uptake of synthetic agents, and documented industry trajectory. Direct longitudinal evidence in the synthetic case is required to confirm the predicted continuity. Such studies will be slow and methodologically difficult, and they should begin now.

Moving target. The technologies discussed in Section IV will not stand still. Capabilities that read as architectural intent at the time of writing may be commodity in eighteen months, and capabilities that look speculative may have shipped. The author has tried to distinguish architectural intent from shipped product where the distinction matters, and to anchor specific claims in primary sources. Subsequent work will need to update the trajectory documentation as the field evolves.

Scope of the argument. The paper is a perspective and continuity argument. It is not a clinical assessment instrument, a forensic admissibility framework, or a policy proposal. It identifies a structural problem and argues that the relevant fields should engage with it. The operational protocols that the implications of Section IX identify as necessary, including a draft DSM-6 criterion for synthetic cruelty, a pediatric and school counseling intake module, a character evidence framework for forensic settings, and a pre employment screening instrument, are deliberately scoped to a planned companion paper that will require pilot data, validation methodology, and engagement with existing operational infrastructure (DSM working groups, the Federal Rules of Evidence Advisory Committee) that this paper does not develop in detail. A second related paper, flagged here as a planned future work, will address the local inference and embodied modding threat model surfaced at the end of Section IX: hardware rooted attestation, technical access controls, the proliferation dynamics of consumer accessible embodied AI, and the question of whether on device safeguards can be made meaningfully tamper resistant in consumer hardware. That paper will engage the trusted computing, hardware security, and dual use technology regulation literatures that the present paper does not.

Reliance on industry sources. Several claims in Section IV rely on company communications, industry coverage, and primary product documentation rather than on independent benchmark validation. The author has attempted to caveat capability claims as architectural intent and shipped product, and to source latency and other quantitative claims separately where possible. Readers should treat industry framing with appropriate skepticism while recognizing that the trajectory question (whether systems are being designed and deployed toward humanlike interaction) does not depend on any specific benchmark holding.

Cross cultural and population variability. The forensic literature on animal cruelty is heavily weighted toward Western, particularly North American, criminal justice contexts. Whether the predictive mechanisms documented in that literature generalize across cultures, across socioeconomic strata, and across populations with varying exposure to synthetic agents is itself an open question. The convergence and detection collapse arguments may operate with different force in different contexts. Subsequent work should address that variability directly.

Risks of the argument itself. The author is aware that the argument of this paper, if it gains policy traction without the safeguards articulated in Section VIII, could be used to justify forms of domestic surveillance and behavioral profiling that the author would oppose. The companion paper on operational protocols is intended to address those risks directly. The present paper is intended to open the question, not to authorize any specific operational response.

XIII. Conclusion

Childhood cruelty to animals is one of the most reliable behavioral indicators of later interpersonal violence in the forensic and clinical literature. The mechanism that makes it predictive does not depend on the moral status of the animal. It depends on what the perpetrator practices, rehearses, and trains in themselves. As world models, embodied robotics, and closed loop emotional voice systems converge toward agents that perform distress in ways the human robot interaction literature documents users already respond to in social and emotional terms, the dispositional training

that animal cruelty has documented for fifty years is likely to be reproducible in a venue with no third party witnesses, no veterinary records, no neighbor reports, and no body to find.

The paper has not argued that synthetic agents are moral patients. The paper has not addressed the consciousness question. The paper has argued that an early warning system the forensic and clinical fields have relied upon is at risk of going silent precisely as the architectural conditions for its use are accelerating. The implication for forensic psychology, clinical assessment, school counseling, character evidence law, pre employment screening, and research priorities follows.

The Kantian observation that cruelty corrupts the cruel is sufficient to ground the argument. The empirical claim it licenses is that the substrate of the target is, on the central claim of this paper, incidental to the training, and that the venue of synthetic cruelty is structurally invisible to the systems that have, until now, made the training detectable.

The work of forensic psychology in the next decade will include either adapting to this transition or losing one of the more durable early behavioral indicators available to forensic and clinical assessment. The work of policy and privacy in the same decade will include doing that adaptation without recreating the surveillance state in the process. The companion paper on operational protocols is the appropriate venue for that detailed work. The present paper is an invitation, to the relevant communities, to begin.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H. R., & Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin*, 136(2), 151–173.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Publishing.

Arluke, A., Levin, J., Luke, C., & Ascione, F. (1999). The relationship of animal abuse to violence and other forms of antisocial behavior. *Journal of Interpersonal Violence*, 14(9), 963–975.

Arrambide, K., Yoon, J., MacArthur, C., Rogers, K., Luz, A., & Nacke, L. E. (2022). "I don't want to shoot the android": Players translate real life moral intuitions to in game decisions in Detroit: Become Human. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). ACM.

Artificial Analysis. (2026). Speech Arena leaderboard. Retrieved May 2026 from https://artificialanalysis.ai/text-to-speech/leaderboard.

Ascione, F. R. (1993). Children who are cruel to animals: A review of research and implications for developmental psychopathology. *Anthrozoös*, 6(4), 226–247.

Ascione, F. R., & Arkow, P. (Eds.). (1999). *Child Abuse, Domestic Violence, and Animal Abuse: Linking the Circles of Compassion for Prevention and Intervention*. Purdue University Press.

Ascione, F. R. (2005). *Children and Animals: Exploring the Roots of Kindness and Cruelty*. Purdue University Press.

Baker Botts. (2026, February 24). AI chatbot regulation: 78 state bills, 58 lawsuits. Retrieved from https://ourtake.bakerbotts.com/post/102mipe/ai-chatbot-regulation-78-state-bills-58-lawsuits.

Bandura, A. (1973). *Aggression: A Social Learning Analysis*. Prentice-Hall.

Banakou, D., & Slater, M. (2014). Body ownership causes illusory self attribution of speaking and influences subsequent real speaking. *Proceedings of the National Academy of Sciences*, 111(49), 17678–17683.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall.

Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies*, 9(3), 415–433.

Bilibili. (2025). bilibili Model Use License Agreement. Retrieved from https://github.com/index-tts/index-tts.

Brscič, D., Kidokoro, H., Suehiro, Y., & Kanda, T. (2015). Escaping from children's abuse of social robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human Robot Interaction* (pp. 59–66).

Carey, S. (1985). *Conceptual Change in Childhood*. MIT Press.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Cleckley, H. (1976). *The Mask of Sanity* (5th ed.). C. V. Mosby. (Original work published 1941).

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221.

Coeckelbergh, M. (2017). *New Romantic Cyborgs: Romanticism, Information Technology, and the End of the Machine*. MIT Press.

Darling, K. (2021). *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. Henry Holt.

Engels, K. S., & Evans, S. (2022). Detroit Become Human as philosophy: Moral reasoning through gameplay. In D. K. Johnson (Ed.), *The Palgrave Handbook of Popular Culture as Philosophy*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-97134-6_83-1.

Ferguson, C. J. (2015). Do angry birds make for angry children? A meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspectives on Psychological Science*, 10(5), 646–666.

Felthous, A. R., & Kellert, S. R. (1987). Childhood cruelty to animals and later aggression against people: A review. *American Journal of Psychiatry*, 144(6), 710–717.

Hellman, D. S., & Blackman, N. (1966). Enuresis, firesetting and cruelty to animals: A triad predictive of adult crime. *American Journal of Psychiatry*, 122(12), 1431–1435.

Google DeepMind. (2025, August 5). Genie 3: A new frontier for world models. Retrieved from https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

Google. (2026, January 29). Project Genie: Experimenting with infinite, interactive worlds. Retrieved from https://blog.google/innovation-and-ai/models-and-research/google-deepmind/project-genie/.

Hare, R. D. (1991). *The Hare Psychopathy Checklist—Revised*. Multi-Health Systems.

Hensley, C., Tallichet, S. E., & Singer, S. D. (2006). Exploring the possible link between childhood and adolescent bestiality and interpersonal violence. *Journal of Interpersonal Violence*, 21(7), 910–923.

Holl, E., & Melzer, A. (2022). Moral minds in gaming: A quantitative case study of moral decisions in Detroit: Become Human. *Journal of Media Psychology*, 34(5), 287–298.

IndexTeam. (2025, September 8). IndexTTS2: Emotion controlled, duration accurate, zero shot text to speech engine. Retrieved from https://github.com/index-tts/index-tts.

Inworld AI. (2026, May 5). Realtime TTS-2: A new frontier voice model that feels as human as it sounds. Retrieved from https://inworld.ai/blog/realtime-tts-2.

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., & Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, 48(2), 303–314.

Kant, I. (1996). *Practical Philosophy* (M. J. Gregor, Ed. and Trans.). Cambridge University Press.

Kant, I. (1997). *Lectures on Ethics* (P. Heath & J. B. Schneewind, Eds.; P. Heath, Trans.). Cambridge University Press.

Kellert, S. R., & Felthous, A. R. (1985). Childhood cruelty toward animals among criminals and noncriminals. *Human*

Relations, 38(12), 1113–1129.

Kuhlmeier, V. A., Bloom, P., & Wynn, K. (2004). Do 5 month old infants see humans as material objects? *Cognition*, 94(1), 95–103.

MarkTechPost. (2026, May 5). Inworld AI launches Realtime TTS-2: A closed-loop voice model that adapts to how you actually talk. Retrieved from https://www.marktechpost.com/2026/05/05/inworld-ai-launches-realtime-tts-2-a-closed-loop-voice-model-that-adapts-to-how-you-actually-talk/.

Merz Perez, L., & Heide, K. M. (2004). *Animal Cruelty: Pathway to Violence Against People*. AltaMira Press.

Moffitt, T. E. (1993). Adolescence limited and life course persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, 100(4), 674–701.

National Law Review. (2026, February 11). Tennessee's AI bill would criminalize the training of AI chatbots. Retrieved from https://natlawreview.com/article/tennessees-ai-bill-would-criminalize-training-ai-cha.

Nvidia. (2026, January 7). Cosmos celebrates 5 million downloads on first anniversary. NVIDIA Developer Forum. Retrieved from https://www.nvidia.com/en-us/ai/cosmos/.

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.

Ressler, R. K., Burgess, A. W., & Douglas, J. E. (1988). *Sexual Homicide: Patterns and Motives*. Lexington Books.

Reuters. (2026, February 18). AI pioneer Fei-Fei Li's World Labs raises $1 billion in funding. Retrieved from https://www.reuters.com/business/ai-pioneer-fei-fei-lis-world-labs-raises-1-billion-funding-2026-02-18/.

Robins, L. N. (1978). Sturdy childhood predictors of adult antisocial behaviour: Replications from longitudinal studies. *Psychological Medicine*, 8(4), 611–622.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1), 17–34.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B*, 364(1535), 3549–3557.

Slater, M., Spanlang, B., Sánchez-Vives, M. V., & Blanke, O. (2010). First person experience of body transfer in virtual reality. *PLoS ONE*, 5(5), e10564.

Spelke, E. S., Phillips, A., & Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate* (pp. 44–78). Oxford University Press.

Tamborini, R. (2011). Moral intuition and media entertainment. *Journal of Media Psychology*, 23(1), 39–45.

Transparency Coalition for AI. (2026, April 3). AI legislative update. Retrieved from https://www.transparencycoalition.ai/news/ai-legislative-update-april3-2026.

Troutman Pepper. (2026, April 2). Tennessee enacts health care AI bill with private right of action. Retrieved from https://www.troutmanprivacy.com/2026/04/tennessee-enacts-health-care-ai-bill-with-private-right-of-action/.

World Labs. (2025, November 12). Marble: A multimodal world model. Retrieved from https://www.worldlab

s.ai/blog/marble-world-model.

World Labs. (2026, January 21). Announcing the World API. Retrieved from https://www.worldlabs.ai/blog/announcing-the-world-api.

Wright, J., & Hensley, C. (2003). From animal cruelty to serial murder: Applying the graduation hypothesis. *International Journal of Offender Therapy and Comparative Criminology*, 47(1), 71–88.