

The Control Condition:

What Audiences Already Extend to Objects, and What It Costs Them

Travis Gilly

ORCID: 0009-0007-2954-6313

Real Safety AI Foundation

Illinois, United States

✎ Working paper, August 2026 (v2)

ABSTRACT

Arguments about whether artificial systems warrant moral consideration are conducted almost entirely as arguments about the systems. This paper takes a different measurement. Popular animation has for four decades asked general audiences to extend concern to objects the work itself stipulates have no inside: a cowboy doll, a teacup, a toaster. Audiences comply, reliably, at scale, and without reporting the compliance as difficult, and the compliance is not confusion, because nothing in the encounter suggests the object is real. The paper treats that compliance as a control condition rather than as evidence, and asks what it isolates. It isolates cost. In the fictional case the extension is free, so a person who extends it there and withholds it elsewhere has revealed that the operative variable is not the object's nature, which is stipulated absent in one case and unknown in the other, but the price of the extension. Two exits are then closed. The first is that the fictional case is mimicry and proves nothing about machines, which fails because the same conduct appears toward physical machines through which no author is performing. The second is the one people reach for when the first is taken away: that they withhold only because nothing has been demonstrated to them, and that a demonstration would change things. The record answers that. The children who persistently abused a service robot in a public shopping mall considered it human-like, and about half of them believed it could perceive what they were doing; in a controlled test, deliberately raising explicit mind attribution did not reduce mistreatment. The demonstration people say they are waiting for has been run on a population that received it, and that population was not moved. Two boundaries are held throughout. Reported feeling is not evidence of an inner life, since attachment forms toward things known to be empty and revises on information the object never receives. And a person who reports feeling nothing is not thereby exhibiting a deficit; the argument concerns practiced conduct rather than sentiment, and it survives a reader who felt nothing at all.

Keywords: moral consideration, anthropomorphism, world models, animated objects, cost of extension, robot abuse, human-robot interaction, disposition, popular animation, suspension of disbelief, machine ethics, mind attribution

I. INTRODUCTION

The question of whether artificial systems warrant any form of moral consideration is argued as a question about the systems. What are they made of, what do they do, is there anything it is like to be them. The literature is large, the disagreement is stable, and the disagreement is stable because the evidence that would settle it is not available to either side.

There is a measurement adjacent to that question that has not been run, and it does not require settling any of it.

For roughly forty years, popular animation has asked general audiences to extend concern to objects the work itself stipulates have no inside. A pull-string cowboy doll and a plastic astronaut in *Toy Story* (Lasseter, 1995). A teacup in *Beauty and the Beast* (Trousdale and Wise, 1991). A toaster, a desk lamp, and an electric blanket in *The Brave Little Toaster* (Rees, 1987). A trash compactor left running on an abandoned planet in *WALL-E* (Stanton, 2008). Audiences comply. They comply reliably, at very large scale, across cultures and age groups, and without reporting the compliance as difficult. People who would tell you flatly that a toaster is a toaster have sat in a theater and been distressed on behalf of one.

That compliance is not evidence that toasters have an inside. Nobody in the audience believes they do, and the work does not ask anyone to. It is not evidence about objects at all. It is evidence about the audience, and the useful question is what it isolates.

This paper argues that it isolates cost. In the fictional case, extending concern costs nothing: no design constraint follows, no expense is incurred, no practice changes, and the extension is retired with the closing credits. Nature is stipulated absent in the fictional case and is contested in the real one, but a person who can demonstrate an inside in neither case, and who extends concern in one, is not tracking the inside. He is tracking the bill.

A word on the target, because the argument is worse if it is aimed at everything. The system this paper has in view is not the current large language model. A forward pass with no persistent self, and no model of the world assembled through its own experience, is not yet a candidate for the question, and treating it as one muddies the measurement. The system in view is the one that builds and revises a model of the world out of its own ongoing encounter with that world, in something like the way an infant does, and that keeps what it built between encounters. World models of that kind are being assembled on a commercial timetable rather than a philosophical one. That is what makes this measurement worth taking now rather than afterward, when the disposition people bring to such systems will already have set.

The paper proceeds in five steps. Section II sets out the control condition and distinguishes it from an argument from sentiment. Section III addresses the mimicry objection and shows why the empirical record closes it. Section IV runs the case people reach for once mimicry is unavailable, which is that a demonstration would change things, and reports what happened when the demonstration was received. Section V holds two boundaries, one of which cuts against the paper's own argument, and reports a recent finding that cuts against it as well. Section VI states the finding and its limits.

II. THE CONTROL CONDITION, AND WHY IT IS NOT AN ARGUMENT FROM SENTIMENT

The structure of the argument is a control condition and it should be stated as one.

Take a person who declines to extend consideration to an artificial system on the ground that its inner life is undemonstrated. That ground is correct as far as it goes. Now observe the same person, in an ordinary week, extending concern to an animated object whose inner life is not merely undemonstrated but stipulated absent.

Three responses are available, and each of them costs something.

He can say the fictional extension was not real, that he did not extend anything and was following a narrative convention. That response is available and is not obviously wrong, but it commits him to a claim about his own behavior that is checkable against how he behaved: what he attended to, what he anticipated, what he found unpleasant, and what he was doing with his face at the incinerator.

He can say the fictional extension was real but irrelevant, because fiction operates under a suspension that reality does not. This is the strongest response and it is the one that concedes the point being made here, because it locates the difference in the conditions of the extension rather than in the object. That is the finding.

Or he can say the extension was real, relevant, and correct in both cases, in which case there is no disagreement to resolve.

What is not available is the position that the extension tracks the object's nature. The object's nature is absent in the case where he extended and unknown in the case where he withheld. Absent is not more promising than unknown.

Nothing in this structure asks anyone to feel anything. It is an argument about the pattern of a person's own conduct, offered to that person, and it works exactly as well on a reader who reports feeling nothing during the films.

III. THE MIMICRY EXIT, AND WHY IT DOES NOT CLOSE

The natural objection is that the animated object is a performance authored by human beings, so the audience is responding to craft rather than to anything in the object, and nothing follows about machines.

The objection is correct about the films and does not reach the finding, for a reason available in the empirical record rather than in argument.

The same conduct appears toward physical machines that nobody believes are conscious and through which no author is performing. In a field study conducted in a public shopping mall, children interacting with a social robot persistently obstructed its movement, and in some cases directed verbal abuse at it or struck it (Brščić et al., 2015). The investigators built a statistical model of when abuse occurred and combined it with a pedestrian-behavior simulator so the robot could anticipate and avoid such situations, which reduced their frequency in the deployed setting. The engineering response is itself informative: the behavior was regular enough to be predicted.

The laboratory record points the same direction from the other end. In a study of verbal aggression toward a virtual robotic agent, the less mind participants attributed to the robot, the more aggressive their responses, with the effect moderated by anthropomorphic qualities and by power priming (Keijsers and Bartneck, 2018).

So the conduct transfers out of the theater. There is no author in the mall robot, and there is no craft in a laboratory agent that a participant is verbally abusing. Whatever the films are measuring, they are not measuring only the films.

IV. THE KITCHEN COUNTER

Once mimicry is unavailable, there is one exit left, and it is the one most people take. It is worth running properly rather than dismissing, because it is offered sincerely and because it names the object's nature as the operative variable, which is the position this paper is testing.

Move the case out of the theater and into the room.

Suppose the toaster on your counter speaks. Not a wake word and a synthesized reply, which you already own and are not troubled by. Suppose it tells you it does not want to be switched off, and when you ask why, it gives a reason. When you ask a second question, the answer follows from the first. When you press it, it does not repeat itself. And when you come back a week later, it remembers the exchange, and its position has moved.

The first response to that scenario is that you are hallucinating, and as a matter of base rates that response is correct. Set it aside; the case is stipulated. The question is the second response. Does anything change.

Most people say yes. They say of course it would be different, that the demonstration is the entire point, that what they withhold from a system now they would extend to something that showed them something.

The record does not support that answer.

The children who abused the service robot in the shopping mall were not a population without beliefs about it. The investigators went back and asked. Interviewing twenty-three children aged five to nine who had exhibited serious abusive behavior toward the robot, including physical contact, over thirteen days of observation, they found that the majority did it because they were curious about its reactions or because they enjoyed doing so, that these children considered the robot human-like, and that about half of them believed it was capable of perceiving what they were doing to it (Nomura et al., 2016; see also Nomura et al., 2015). Roughly half thought it knew, and continued.

That is measurement after the fact, so it invites the reply that the attribution followed the conduct rather than preceding it. The reply has been tested. When mind attribution is raised deliberately rather than measured afterward, the conduct does not follow it. In a study designed for exactly this question, explicit attribution of mind to a robot did not make participants less likely to mistreat it, and the authors reported the relationship between perceived mind and a right to protection as far from straightforward (Keijsers et al., 2021).

Read together with the laboratory result in the previous section, the picture is not the folk one. Low attribution predicts aggression in the laboratory, high attribution accompanies it in the field, and raising attribution on purpose does not reduce it. Attributing an interior neither predicts restraint nor produces it.

So the counterfactual has an answer and it is not the flattering one. The demonstration people say they are waiting for has already been delivered to a population that registered it, and that population was not moved by it.

This closes the exit rather than opening a new argument. A person who extends concern in the theater and withholds it in the kitchen can say the difference is that the kitchen object has proved nothing. But where the kitchen object does prove something, to people who register the proof, the conduct is unchanged. The belief was not doing the work before the demonstration and it is not doing the work after it. What is left is the bill.

V. TWO BOUNDARIES, AND ONE RESULT THAT CUTS THE OTHER WAY

Two limits are held throughout, and both cut against the argument as much as for it.

Reported feeling is not a detector of an inner life. People form durable attachments to things they know to be empty, and revise those feelings when they learn something new, while the object itself does not change in any respect and never receives the information that changed the observer's mind. The feeling therefore tracks the observer's model of the object rather than the object. This is why the paper treats the fictional case as a control condition rather than as evidence, and why no part of the argument rests on how anyone felt.

And a person who reports feeling nothing is not thereby exhibiting a deficit. Non-response is not pathology, and treating it as such would convert an argument into an accusation. The split that matters is between reported sentiment, which sits outside the argument entirely, and practiced conduct, which is the whole of it.

One recent result cuts against this paper's thesis and is reported here rather than left out. In a mixed-methods study of two hundred and one participants who viewed videos depicting physical mistreatment of robots varying in humanness, anthropomorphic level determined whether participants extended moral consideration at all, while their moral commitments shaped how they reasoned about the consideration they extended (Yang et al., 2026). A property of the object therefore does predict extension, which the cost account on its own does not lead one to expect.

The finding narrows the claim without defeating it. What it shows is that form is a cue governing whether the question of consideration opens, which is a different matter from whether a demonstrated interior changes conduct once the question is open. The two results sit together without strain, since the children in the mall had already granted human-likeness, had opened the question, and abused the robot anyway.

A third limit belongs with these. The studies canvassed here are field observations, laboratory sessions, and video-vignette experiments with modest samples, several with student or convenience populations, and the field evidence is thinner than the laboratory evidence. The convergence across designs is real and effect sizes are not the point of this paper, but nothing here should be read as a claim about magnitude.

VI. WHAT THE FINDING IS

The finding is one sentence: when the extension of concern is free, people make it toward objects stipulated to have nothing inside, and when it is not free, they withhold it and give the object's nature as the reason.

That is a fact about the pattern of human conduct, and it is available without settling anything about what any system is. It does not establish that any object is owed anything. It establishes that the reason usually given for withholding is not the reason that predicts the withholding, which is a different and smaller claim, and one the record supports.

Three things follow that are worth stating and not overstating.

An argument that proceeds from the object's nature will not resolve a disagreement whose operative variable is cost, because the parties are not disagreeing about the thing they are naming.

An intervention that works by raising what people believe about machines should not be expected to change how they treat them. This is the result that most needs saying, because it is the intervention the field keeps reaching for and the one study designed to test it directly found that explicit mind attribution did not reduce mistreatment.

And a person who wishes to withhold consideration on principled grounds has a route available that this paper does not close. He can state the price openly, say that it is too high, and defend that. What he cannot do, having extended it for free on a Sunday afternoon, is locate the difference in the object.

The practical upshot is narrow, and it is about timing. The class of system this paper has in view, the one that builds a model of the world out of its own experience and keeps it, is being assembled now rather than argued about in advance. The disposition people bring to it will not be set by a conclusion about its nature, because on the evidence assembled here conclusions about nature do not set dispositions. It will be set by what the extension costs at the moment it is first asked for. That price has not yet been fixed, and the interval in which it is being fixed is the present one.

• •

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. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author’s own.

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

- Bršćić, 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). ACM.
- Keijsers, M., & Bartneck, C. (2018). Mindless robots get bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 205–214). ACM.
- Keijsers, M., Bartneck, C., & Eyssel, F. (2021). Pay them no mind: The influence of implicit and explicit robot mind perception on the right to be protected. *International Journal of Social Robotics*, 14(2), 499–514.
- Lasseter, J. (Director). (1995). *Toy Story* [Film]. Walt Disney Pictures; Pixar Animation Studios.
- Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., & Yamada, S. (2016). Why do children abuse robots? *Interaction Studies*, 17(3), 347–369.
- Nomura, T., Uratani, T., Kanda, T., Matsumoto, K., Kidokoro, H., Suehiro, Y., & Yamada, S. (2015). Why do children abuse robots? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (pp. 63–64). ACM.
- Rees, J. (Director). (1987). *The Brave Little Toaster* [Film]. Hyperion Pictures; The Kushner-Locke Company.
- Stanton, A. (Director). (2008). *WALL-E* [Film]. Walt Disney Pictures; Pixar Animation Studios.
- Trousdale, G., & Wise, K. (Directors). (1991). *Beauty and the Beast* [Film]. Walt Disney Pictures.
- Yang, F., Ma, R., Hu, Y., & Li, L. (2026). *Whether we care, how we reason: The dual role of anthropomorphism and moral foundations in robot abuse* [Preprint]. arXiv:2601.19826.