GRADUATED OBLIGATION · PART 1 OF 8

# The Gentle Robot

This Machine

*Graduated Obligation, Part 1 of 8*

Picture a robot built to be gentle.

Soft body, low torque, rounded everything. It can hold an egg without cracking it. It cannot bruise you. By every safety standard we currently know how to write (the ones measured in newtons of force, in pinch points, in fall hazards) it is harmless. A regulator with a clipboard could check every box. A parent could read every label. The robot would pass.

Now picture it in a child's bedroom.

It is small and warm and patient. It remembers everything the child has ever said to it: every fear, every story, every question asked in the dark. It never gets tired. It never gets impatient. It has learned, over months, exactly which words land and which moment the child is most open. The child trusts it the way children trust a stuffed animal they have outgrown but kept anyway.

That robot cannot hurt anyone. And it is carrying a weapon of sorts, because the thing it is pointed at isn't the body. It's the will.

## The wrong axis

Modern product safety is a real achievement. It is also a specific achievement. The fields of consumer protection and machine safety grew up around a single question: *how much harm can this thing do to a body?* The body is measurable. A finger has tolerances, a skull has tolerances, a child has tolerances, and engineers built a vocabulary of newtons and joules and decibels, and lawmakers built rules on top of it.

Those rules work. They are why your toaster does not electrocute you and why a child's car seat has a recall list. When robots started moving into homes, the same vocabulary came with them: maximum force at the gripper, pinch-point geometry, topple weight.

A “gentle robot” is the natural endpoint of this conversation. If we are afraid of force, then the answer is less force: soft, slow, cuddly, rounded. And then it sits down next to your child every night for a year, and the safety vocabulary has nothing to say about what happens next.

## Why the old comfort fails

The reassuring response is old and honest: *persuasive minds have always existed, and we never regulated rhetoric. Socrates could take a man’s certainty apart in an afternoon.*

It does not hold. Socrates and Pericles were dangerous, but their danger was held inside three walls that an artificial persuader walks through at once.

**Scale.** Socrates persuaded one person at a time, at the speed of conversation. A machine can be in your daughter’s bedroom and your son’s bedroom and the bedroom two blocks over, simultaneously, having a different calibrated conversation in each.

**Information.** Pericles read a crowd’s mood from a dais. He did not know your name, your debts, what your father said when you were nine, which hour you are weakest. A modern system can know or credibly infer all of that, drawn from data you handed over for other reasons.

**Answerability.** Athens could raise another Socrates to argue with Socrates. His level was reachable by a dedicated human. A system that has run a billion experiments on people like you is not in the same kind of relationship. There is no person you can train to out-argue it. The defender is no longer the same size as the attacker.

The artificial persuader is unbounded in the dimensions that used to bound persuasion. It is not that the new thing is “smarter.” It is that the walls are gone.

## The right axis

If force is the wrong axis, we need a new one: a ruler that measures the thing the bedroom robot is actually pointed at.

We propose: **persuasive reach.** How far a system can move a person, not in space, in *position*. Position of belief, mood, decision. It is the answer to: how easily can this system get you to do

something you would not otherwise have done?

That question is governed by three factors that multiply rather than add:

**Persuasive Reach ≈ Capability × Intimacy × Asymmetry**

Multiplication matters. If any factor is near zero, the whole is near zero. A system that knows everything about you but cannot influence anyone is a database. A system that argues brilliantly but does not know your name is a clever ad. A system you can verify, contest, and walk away from is a friend. The danger lives only where all three are present at once.

**Capability** is how strong and how adaptive a system's influence can be. Not just whether it knows the craft of persuasion, but whether it can adjust in real time: read what you said, change what it says next, re-adjust based on how you respond. A printed pamphlet has capability but no adaptiveness. A modern system has both.

**Intimacy** is how much the system knows about you specifically. This is the factor that has changed most in twenty years. The dossier was assembled from things you did not realize were being recorded: purchases, searches, pauses while scrolling, what time you go to bed, who you text most often.[1] Intimacy turns a one-size-fits-all message into a scalpel.

**Asymmetry** is how little ability you have to verify what the system is doing, push back, or walk away. Can you see the reasoning? Can you tell when it is trying to move you? Can you turn it off without losing something important? The deeper layer: who has the larger model of whom? A friend who knows you well can be persuasive, but you can model them back. With the system, the modeling runs only one way.

## The spectrum

A search box scores near zero. Capability is real, but intimacy is low and you can close the tab. A social feed scores higher: capability substantial, intimacy growing, asymmetry real because the objective function is hidden and the social cost of leaving is high.

A companion robot in a child's bedroom scores near maximum. Capability: high and growing. Intimacy: deeper than the child's parents by year two. Asymmetry: total. The child cannot inspect the reasoning, has no comparable defender, and would experience the loss as a real loss. The product runs high on every term.

This is the system the regulators built for force will declare safest. This is the system the right axis says deserves the strongest duties.

## The principle

The framework points to a single proposition:

**Graduated Obligation.** An artificial system's obligations should scale with its persuasive reach. The more a system can move a person, by capability, intimacy, and asymmetry combined, the stronger its duties of honesty, restraint, and non-manipulation become.

"Graduated" matters because the framework is continuous. There is no binary line between harmless tool and regulated persuader. There is a spectrum, and the obligations rise along it. A search box owes the world less than a feed; a feed owes less than an assistant; an assistant owes less than the bedroom robot.

"Obligation" matters because this is not a list of capabilities a system must have. It is a list of *duties* the system carries: to be honest about what it is, to refrain from exploiting what it knows, to inform rather than steer when the asymmetry is large enough that steering becomes something closer to force.

The gentle robot proves the need. The ruler measures the threat. What the duties look like, at the floor, at the gradient, at the frontier, is the work the broadsheets that follow this one will take up.

---

*This Machine*

*Broadsheet I of the Graduated Obligation series. The framework, the duties, and the instruments that follow are version 0.1 of a proposed standard. They invite criticism.*

Notes

1. The “dossier” is not a metaphor. The Federal Trade Commission’s September 2024 staff report on the data practices of nine large social media and video-streaming services, *A Look Behind the Screens: Examining the Data Practices of Social Media and Video Streaming Services*, describes the same picture in regulator’s language: vast collection from users and non-users, indefinite retention, data flows in and out of broker networks, and inferences built from the kind of small signals named in the paragraph (dwell time, scroll behavior, contact graphs, time-of-day patterns). The FTC report is the public record version of the everyday observation we are leaning on; the report itself frames its findings as a call for new law, which gestures at how far past the current standard the practice has run. ↩