Responsibility and Accountability

Table of contents

Mark Coeckelbergh’s paper “Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability” (Science and Engineering Ethics, 26:2051–2068, 2020).

---

1. The Core Problem: Responsibility for AI

Argument Overview
Coeckelbergh starts by stating that artificial intelligence (AI)—particularly machine learning systems—creates pressing concerns about who should be held accountable for *good* and *bad* outcomes once decisions are automated. He notes the “urgent” character of these questions, citing examples like self-driving cars, automated financial trading, or Boeing’s autopilot systems, which lead to real and sometimes tragic consequences. The key ethical puzzle is:

> “Given that AI enables society to automate more tasks, who or what is responsible for the benefits and harms?” cite turn0file0

Two Conditions from Aristotle
He uses *Aristotelian* criteria for responsibility—**(1) control** and **(2) knowledge**—to anchor the discussion. Traditionally, we say a person is responsible for an action if:

1. They cause it voluntarily or freely (the *control* condition).
2. They are not ignorant of what they are doing (the *knowledge* condition).

Coeckelbergh takes these two conditions (which he calls “Aristotelian” or “standard”) as the platform to show how AI’s complex and opaque nature complicates both.

Transition to the Puzzle
On the one hand, software can “do” things that look agent-like. On the other hand, we are not ready—legally or philosophically—to grant machines *moral agency*. This sets up the question: **If AI is not a responsible agent, how do we assign responsibility to the humans involved?** That question spans from straightforward accidents (e.g., an autonomous Uber hitting a pedestrian) to subtler issues of bias in data sets or mass surveillance.

---

2. Only Humans Are (Still) Moral Agents—But That Doesn’t Solve the Attribution Problem

Machines Are Not (Yet) Responsible Agents
While acknowledging the debate on whether AI might ever *qualify* as a full moral agent, Coeckelbergh assumes:

> “AI technologies can have agency but do not meet traditional criteria for moral agency and moral responsibility.” □cite□turn0file0□

Hence, only humans can bear responsibility. Machines lack consciousness or freedom in the strong sense.

The “Many Hands” Problem
Yet even if we insist responsibility *must* rest with humans, there remains a thorny issue: advanced AI applications typically involve a large network of people: designers, coders, managers, corporate owners, end users, regulators, data providers, and so on. Coeckelbergh references the *problem of many hands*:

> “Who is responsible? It could be the developers of the software, the car company, the user, the regulator ... and within the category ‘software development’ there may be a lot of people involved.” □cite□turn0file0□

In short, so many individuals play a part that assigning moral or legal liability to one or two specific people is incredibly difficult. This is made even more complicated by:

- **Temporal gaps:** The code or dataset might have been produced long ago by someone no longer reachable.
- **Geographic dispersal:** Different teams around the world might each contribute to small pieces of the project.

“Many Things” Complicate It Further
Coeckelbergh adds that it is not *only* “many hands”; it is also “many *things*.” AI technology is layered on top of sensors, other pieces of software, mechanical subsystems, or user interfaces. A malfunction in a sensor can cause an accident that we might hastily blame on the AI. Meanwhile, a glitch in the dataset pipeline could introduce bias. As he puts it:

> “One of the problems with technological action is that there are usually many people causally involved ... but also many devices, programs, and interacting parts.” □cite□turn0file0□

Hence, the “black box” of AI is part of an *even bigger* network of machinery and code. This means effective responsibility-attribution demands understanding every layer of software/hardware, plus how they interact.

---

3. The Knowledge Condition: Transparency, Epistemic Gaps, and “Explainable AI”

Aristotle’s Second Criterion: No Ignorance
Following Aristotle’s *Nicomachean Ethics*, Coeckelbergh says an agent must not be ignorant of what they are doing:

> “A man may be ignorant ... of who he is, what he is doing, what or whom he is acting on ... and sometimes what (instrument) he is doing it with.” □cite□turn0file0□

To adapt this to AI, a software engineer (or a judge using an AI tool) might *think* they know how the program works, but because machine-learning systems can be so opaque, they lack full clarity on *why* the system reaches specific decisions. Even developers can be “ignorant” of the machine’s internal reasoning, especially in deep learning systems.

Examples of “Black Box” Systems
Machine learning, especially deep neural networks, can produce outputs that surprise even their creators. A so-called *self-driving car* may not be able to articulate the chain of reasoning behind, say, turning left too late. Likewise, a judge or parole officer using AI-based “risk scoring” might not grasp *exactly* how the system weighted various factors. Coeckelbergh frames it this way:

> “[The user] suffers from the ignorance of not sufficiently knowing their instrument … they do not know what they do when they give a recommendation to someone based on this kind of AI.” □cite□turn0file0□

Thus, there is a worry that the more complicated or “black box” the technology, the less users can meet the knowledge condition for moral responsibility. This sets up one rationale for “Explainable AI” (often shortened to “XAI” in other literature).

---

4. A Relational Take on Responsibility: Agents *and* Patients

Going Beyond Agency
Standard debates often focus only on *agents* (those who do the action) and whether they have control or knowledge. Coeckelbergh argues that we forget the other side: the moral *patients* who are affected, harmed, or otherwise impacted by AI-driven outcomes. In a more *relational* framework:

> “We should not neglect the problem of the addressee. Those to whom moral agents are responsible … who demand reasons for actions and decisions made by using AI.” □cite□turn0file0□

Here, he draws on the idea of “answerability.” Responsibility does not merely mean you *have* knowledge but that you must be prepared to *give* an account (to actual people impacted by your action).

Explainability as Answerability
Coeckelbergh’s *fresh twist* is that it is not enough that a human developer or user personally understands the system’s output; they need to be able to *explain it to others*. A credit applicant denied a loan, a parole candidate kept in prison, or a passenger on an airplane that goes off-course is *owed* an explanation from the responsible humans. Hence:

> “Explainability is not only a matter of knowledge on the part of the agent. … The agent needs to be able to explain to the patient why she does or did a particular action.” □cite□turn0file0□

This *relational* dimension underscores the importance of designing AI systems and organizational processes so that an official—*not* the machine alone—can step forward to clarify (to real people) why a particular decision was reached.

Collective, Distributed, or Shared Responsibility
This relational approach also supports the idea that multiple human agents, collectively, might shoulder responsibility. Furthermore, if a dataset is biased because *society* itself has longstanding discriminatory patterns, the blame might not lie with a single programmer but with a broader social collective. Coeckelbergh even calls this possibility “tragic,” in that no single party can unravel centuries of cultural bias. Still, those deploying AI cannot simply *excuse* themselves: they have a duty to mitigate or correct biases, or at minimum to *explain* them.

---

5. Concrete Examples

Throughout, Coeckelbergh applies his analysis to real and hypothetical scenarios, such as:

1. **Self-Driving Cars**:

    - If an accident occurs, we might blame the software developers, the sensor manufacturers, the occupant who might have “dozed off,” or the city for not having proper road markings. This is the epitome of “many hands” plus “many things.”
    - Transparency is key: a good system should let an investigator reconstruct what signals the AI was processing and how the occupant was meant to oversee it.

2. **Boeing 737 MAX Crashes**:

    - The autopilot software repeatedly pushed the nose down. Pilots seemingly lacked time or full knowledge to override. Coeckelbergh highlights that with advanced automation, humans may not have enough time to intervene. This time pressure underscores the loss of *control*, raising the question: “How can we hold them responsible if they could not feasibly intervene?”

3. **Military Autonomous Systems**:

    - Fully automated missile defense or lethal autonomous weapons make real-time human oversight impossible. This scenario intensifies the moral stakes. He warns that if we keep building such systems, we create a responsibility gap in which nobody can meaningfully control or even know exactly what the system is doing at lightning speed.

4. **Bias in Machine Learning**:

    - Data reflecting sexist or racist patterns might produce discriminatory outputs. Who is responsible? The original data annotators? The entire society that used biased language? The developer who never tested for bias?

These examples unify Coeckelbergh’s broader message about the complexities of moral blame in large-scale, multi-actor, multi-layered AI systems.

---

6. Explainability Techniques and Policies

Technical Solutions
Coeckelbergh mentions the emerging field of *Explainable AI*, referencing methods like heatmaps or local interpretable model-agnostic explanations (LIME), though he does not dive deeply into their specifics. He does, however, stress that:

> “Technical ‘explainability’ ... should be seen as something in the service of the more general ethical requirement of explainability and answerability on the part of the human agent.” □cite□turn0file0□

Legal or Regulatory Measures
He also touches on how GDPR in Europe gives people a right to certain kinds of information but does *not* necessarily give a “right to an in-depth explanation.” Policymakers might need to extend these protections so that impacted individuals can ask “Why?” and actually get a reasoned answer. In his view, we may need:

- *Traceability* rules that require data provenance and logging.
- *Impact assessments* or “checkpoints” to ensure that an AI’s complexity does not completely mask the chain of responsibility.

---

7. The Tragic Dimension: Limits of Agency and Collective Action

Coeckelbergh acknowledges that even the best frameworks may run up against historical, cultural, or structural injustices. Suppose an entire language and society exhibit deep biases that seep into text corpora. Suppose technology evolves so that no single actor can possibly reconstruct everything:

> “Responsibility for AI and other technologies may be limited to some degree and has a tragic aspect.” □cite□turn0file0□

He connects “tragedy” to the idea that there are sometimes no *perfect* solutions. *However*, he does not see that as a free pass to shirk moral reflection. Rather, individuals and institutions should do their part to improve transparency, anticipate biases, and remain accountable—even if they cannot guarantee the total elimination of harm.

---

8. Conclusion: Relational Responsibility as the Way Forward

From “Control and Knowledge” to “Answerability and Patiency”
Coeckelbergh’s key conclusion is that moral and legal discussions of AI responsibility must:

1. *Re-assert* that *humans* must remain the ultimate bearers of responsibility—even if AI looks increasingly autonomous.
2. *Acknowledge* that control is increasingly complicated by quick operations and complex networks, so we need new tools (technical, legal, organizational) to keep track of who does what.
3. *Emphasize* knowledge or “explainability” not only for the sake of the agent’s own clarity but also for everyone affected. Responsibility is a *relational* practice in which impacted parties deserve an explanation.

He underlines that “responsibility as answerability” surpasses purely theoretical conceptions of control or intent, by foregrounding the *human-to-human* requirement that reasons or explanations be given:

> “In the end, only humans can really explain and should explain what they decide and do.” □cite□turn0file0□

Practical Prescriptions
As a result, Coeckelbergh calls for:

- **Developing AI** in ways that support *human* moral agency (including giving humans enough time or access to intervene).
- **Improving traceability** so that investigators and stakeholders can see how decisions were formed.
- **Embedding AI** in contexts (courts, corporate offices, legislative frameworks) that demand answerability so that the “patients” can challenge or question a machine-driven decision.

---

9. Key Takeaways and Final Reflection

1. **No Simple Answers**: There is no single “magic bullet” for distributing responsibility in large socio-technical systems.
2. **Control + Knowledge**: The classical conditions matter, but they are *strained* by AI’s complexity and speed.
3. **Relational Responsibility**: The author’s unique contribution is highlighting that responsibility is *not* only about the agent’s moral agency but about how that agent *answers to* real people impacted by the AI’s decisions.
4. **Collective and Cultural Dimensions**: The problem might be partly *structural*, requiring collective reforms of data practices and language.

In short, Coeckelbergh’s text systematically shows how AI’s complexity complicates responsibility while insisting we must still locate accountability in *humans*. He enriches the debate by reminding us that “explainability” is not just a technical feature but a moral demand that arises in a social relation between those who make decisions (or build the deciding machines) and those whose lives are shaped by them.

References in the Text

- *Aristotle, Nicomachean Ethics*
- *Bostrom (on superintelligence)*
- *Taddeo & Floridi (“distributed responsibility”)*
- *Matthias (2004), “The Responsibility Gap”*
- *Floridi & Sanders, Moor, Gunkel (discussions on moral agency and robotics)*
- *Levinas (on the face of the other)*
- *Johnson, Sparrow, Wallach & Allen (machine ethics)*
- *Latour-inspired “actor-network” conceptions (Hanson)*
- *Caliskan et al. (2017) on bias in language corpora)*
- *Miller (2019) on social aspects of explanation*
- *Sunstein, Kleinberg, etc., about human versus algorithmic biases*

Coeckelbergh cites these and more to demonstrate that AI ethics is interdisciplinary—drawing from philosophy of technology, legal theory, moral psychology, and broader social theory.

Final Word

This paper thus provides a *philosophically grounded, yet practically urgent* argument for viewing AI responsibility in light of classical ideas about moral control and knowledge, while *updating* them for large, fast, and opaque socio-technical systems. Coeckelbergh’s **relational justification of explainability** is his most distinctive insight: ultimately, if we cannot supply reasons or clarifications to those impacted by AI, *we* (the humans) have failed to exercise our moral responsibility—no matter how advanced the machine. □cite□turn0file0□

Stanford Encyclopedia of Philosophy (SEP) entry “Computing and Moral Responsibility” (updated Thu Feb 2, 2023)

---

1. Introductory Framework and Key Questions

Right at the start, the article notes that **traditional accounts of moral responsibility** typically focus on *human* actors performing *direct* actions with *visible* outcomes. However, it stresses that **in the modern, technologically driven world,** humans engage with a vast network of *sociotechnical* factors. Computing technology:

- Shapes our decisions,
- Facilitates or constrains our actions,
- And thus complicates the classical framework of attributing responsibility.

The authors explicitly connect these points to Jonas (1984), pointing out that conventional moral frameworks were never designed to handle the "reach" and complexity of modern computing, nor the ways technology "actively mediates" actions (citing Latour 1992; Verbeek 2021). The main question they pose is:

> "Are human beings still fully responsible for the effects produced by technologies that they do not (and perhaps cannot) fully understand or control?"

This question undergirds the rest of the entry.

---

2. The Three Traditional Conditions for Moral Responsibility

To show how computing complicates moral responsibility, the article isolates **three conditions** typically invoked in Western moral philosophy when attributing responsibility (drawing on Jonas 1984, among others):

1. **Causal Contribution:** The person or group must have some control or causal influence over the outcome.
2. **Foresight / Knowledge:** The person must be able to *anticipate or consider* the consequences of their actions.
3. **Freedom or Voluntariness:** The person's action must be *sufficiently free,* not coerced by forces beyond their control.

Though these conditions are often contested in philosophy (e.g., debates around free will or knowledge), the SEP entry highlights that **computing** raises new, *intensified* challenges for each.

2.1 Causal Contribution (The "Many Hands" Problem)

An especially notorious issue is *the problem of many hands:*

- When modern computer systems fail or cause harm (e.g., the Therac-25 overdoses, or Boeing 737 MAX crashes), *multiple* developers, technicians, managers, regulators, and possibly end users share partial responsibility.
- Tracing direct cause-and-effect to a single "culprit" is nearly impossible.
- The complexity of these "sociotechnical systems" often yields *collective* or *distributed* responsibility.

Additionally, there's *temporal and geographical distance:*

- A programmer might be halfway around the world, or the system may run *years* after development.

- This distance blurs lines of accountability: it is easy for participants to say “I only did a small part,” or “It wasn’t *my* responsibility.”

Hence, the *causal link* is obscured by layering, networks, and time-lag.

2.2 Knowledge or Considering the Consequences

The second condition is that the actor must *realize or foresee* the consequences of their action:

- On the one hand, computers can aid in analyzing data and thereby *improve* our knowledge.
- On the other, they often hide or mask logic “behind the interface,” especially with machine learning. Judges relying on opaque “risk assessment” tools may not know exactly *why* a defendant got a certain score. Similarly, remote pilots of drones or operators of any AI-driven system may not fully appreciate *who* is affected.

The article gives examples like:

- **Automation bias:** People sometimes *over-trust* the computer (as in the USS Vincennes incident, which shot down a civilian aircraft after misidentifying it); or
- **Overwhelming false alarms:** People *under-trust* the system, ignoring genuine signals when they come.

Thus, even when technology can *help* us gather information, it might mislead or systematically hamper understanding. The frequent novelty of technology—e.g., new modes of software-based wrongdoing—can also create moral “grey zones” in which no established norms yet apply, making it unclear how to weigh responsibilities.

2.3 Freedom to Act

A final condition is that individuals must *freely choose* an action, not be coerced. The article underscores:

- **Automation** often removes or reduces *human discretion*. When a process is fully automated (such as a camera automatically issuing speeding tickets), who has freedom? The official? The user?
- Some systems are *explicitly* designed to limit freedom (e.g., an anti-alcohol lock that prevents a car from starting if the driver fails a breath test). Or more subtle “nudges” in user interfaces can shape user behavior—dark patterns that manipulate choices without individuals realizing.

Critics worry that advanced, data-driven “nudges” or “hyper-nudges” (Yeung 2017) *undermine human autonomy*. As systems become more pervasive, the scope for truly independent, reflective choice shrinks.

---

3. Can Computers (or Robots) Be Moral Agents?

A key question: **should we label the technology itself as morally responsible**—or “morally accountable”—when its behavior seems autonomous?

The article lays out multiple arguments:

1. **Computers as Morally Responsible Agents?**

- *Dennett (1997)* proposed that if a system exhibits "higher-order intentionality"—the ability to have beliefs about beliefs—it might be morally responsible.
- *Sullins (2006)* claims that if a robot displays sufficient autonomy, intentional behavior, and a social role, it can be "morally responsible."

However, these positions face pushback, e.g., critics say that computers lack consciousness, real intentionality, and the ability to *suffer* or *be punished* (Sparrow 2007).

2. **Designing 'Autonomous Moral Agents' (AMAs)**

   - *Allen and Wallach* (2012) want to embed moral decision-making capacities in AI (driverless cars or military robots forced to weigh moral trade-offs).
   - *Moor (2006)* differentiates implicit ethical agents (computers that have built-in moral constraints) from explicit ethical agents (which can reason over moral rules) and "full" ethical agents (functionally akin to humans).

   Proponents claim we need such "machine ethics" for advanced systems. Critics respond that even if it's *functionally* helpful to embed moral logic in machines, that does not necessarily grant them full moral responsibility (since humans still design and deploy them).

3. **Expanding the Concept of Moral Agency**

   - *Floridi and Sanders (2004)* suggest that we might treat certain highly interactive, adaptive systems as "morally accountable" *without* holding them morally *responsible* in the traditional sense (they draw an analogy to how we treat animals).
   - Critics (e.g., *Johnson 2006*) say labeling the computer an "agent" deflects moral scrutiny from *humans*—the designers, owners, or users—who ultimately embed intentions in these artifacts.

A recurring theme: seeing the technology as *co-responsible* might obscure or reduce emphasis on the humans "behind" the machine. Authors like *Verbeek* and *Johnson & Powers* argue that moral agency is "hybrid" or "distributed," spanning multiple humans + artifacts. But we must not lose sight of *human choices* that shape those artifacts.

---

4. Rethinking the Concept of Moral Responsibility

Given these challenges, various approaches have been proposed to refine *how* we see responsibility in the computing domain:

4.1 Assigning Responsibility (Positive vs. Negative)

- **Misconceptions about "Ethical Neutrality"**
  Gotterbarn (2001) describes how many computing professionals erroneously regard their field as ethically "neutral"—they see themselves as mere coders fulfilling a specification. But as the article recounts, design choices *always* have moral consequences (even if unrecognized).

- **Malpractice Model vs. Positive Responsibility**
  Gotterbarn notes that a *malpractice* mindset—where responsibility only comes into play if something goes wrong—often leads people to disclaim liability, point to others, or blame complexity. Instead, he advocates *positive responsibility*, meaning professionals proactively anticipate harmful outcomes and

act to prevent them. This is a forward-looking, *virtue-like* stance: you consider your moral duty to make robust, safe code, *regardless* of whether you can be blamed later.

- **Limits to Positive Responsibility**
  The article then addresses legitimate worries: how far can a developer realistically anticipate future misuse or unexpected interactions? Some authors (like *Martin 2019*) say that if a company voluntarily decides to market a system in a certain domain, it assumes an obligation to thoroughly consider *that domain's values,* side effects, biases, or vulnerabilities. That said, the entire chain of actors—designers, managers, corporate owners—must also *share* in that forward-looking responsibility (Santonio de Sio & Meccaci 2020).

4.2 Meaningful Human Control

One explicit approach is **designing sociotechnical systems for "meaningful human control"**:

- *Santonio de Sio & van den Hoven (2018)* argue that to ensure humans remain accountable, we must build systems that:

  1. **Track** moral reasons and relevant facts in a situation (so the system's behavior aligns with the reasons or values of human stakeholders);
  2. **Trace** outcomes back to identifiable human decisions (so there is at least one *informed* human in the loop).

- This approach aims to fix "responsibility gaps" in highly automated scenarios like autonomous weapons. If we can't see *who* or *what* decided to shoot or accelerate, there is no meaningful moral control. By carefully engineering both technology and organizational processes, we can preserve the possibility of attributing moral responsibility.

4.3 Responsibility as a Social Practice and "Culture of Accountability"

Another perspective sees responsibility as *fundamentally interpersonal*, involving "practices of holding others to account." The SEP entry references *Nissenbaum's* argument (1994, 1997) that organizational and social structures sometimes:

- Let people "blame the computer" or accept "bugs" as inevitable, thus *eroding accountability*.
- Shift blame around in large organizations so that no single party invests effort into safer design or usage.
- Issue disclaimers/extended licenses to disclaim liability—"ownership without accountability."

Creating a **culture of accountability** means ensuring that at *every step* (from design to deployment) there is a clear sense that real people must "answer for" the system's performance. That fosters better attention to reliability, potential negative impacts, and ways to remedy or mitigate harm.

Moral Crumple Zones (Elish 2019) are also discussed: the phenomenon where blame unfairly lands on the nearest human operator. This might happen in an aircraft cockpit or with a self-driving car "safety driver," even if they had little real control. If society reflexively puts all blame on these individuals, we neither fix the root cause nor hold the correct parties accountable. The article warns that to get the distribution of responsibility correct, we must look carefully at the entire network of actors and artifacts.

5. Conclusion: Toward a Hybrid, Sociotechnical View of Responsibility

In its final section, the entry reiterates that computing technologies:

- Often disrupt causal clarity (who caused what?),
- Expand or obscure knowledge (who foresaw what?), and
- Restrict or reshape freedom (who actually decides what happens?).

While some authors want to ascribe partial agency to AI, most caution against letting this overshadow the fact that **human** designers, managers, and operators embed their intentions, biases, and constraints into technological artifacts. The recommended path forward thus typically blends:

1. **Collective or Distributed Responsibility**: Recognizing that design and usage of technology is rarely an individual affair.
2. **Positive, Forward-Looking Responsibilities**: Professionals, companies, and institutions should anticipate harms, not merely avoid blame after the fact.
3. **Organizational Measures**: Creating *cultures of accountability* and ensuring that "meaningful human control" remains possible.
4. **Ethical Reflection + Technical Design**: Considering how to embed moral reasoning in code (machine ethics) or, more expansively, how to structure the human-artifact system so that moral reflection and responsibility remain intact.

The article underscores that "any reflection on these concepts (responsibility, autonomy, moral agency) will need to address how technologies affect human action, and where responsibility for action begins and ends." In simpler terms, the SEP entry calls for an expanded, socially aware, and *technically informed* moral philosophy—one that grapples with the complexity of modern computing rather than ignoring it.

Key References Mentioned

1. **Therac-25** and **USS Vincennes** as classic cautionary tales about over-reliance on automated systems.
2. **Jonas (1984)** for the shift in moral philosophy required by modern technology.
3. **Latour (1992), Verbeek (2021)** for the "active mediation" role of technology in shaping human decisions.
4. **Floridi & Sanders (2004)** for treating artificial agents as "accountable" sources of moral action, somewhat analogously to animals.
5. **Moor (2006)** on implicit vs. explicit vs. "full" ethical agents.
6. **Nissenbaum (1997)** on the erosion of accountability, "computer as scapegoat," and the need to foster a culture of accountability.
7. **Gotterbarn (2001)** on misconceptions of "ethical neutrality" and the problems with a purely malpractice model.
8. **Santonio de Sio & van den Hoven (2018)** on meaningful human control, plus elaborations by Santonio de Sio & Meccaci (2020).
9. **Elish (2019)** on "moral crumple zones," the risk that humans near advanced systems end up absorbing blame.

Closing Reflection

Overall, the SEP entry provides a *comprehensive philosophical map* of how computing troubles standard assumptions about moral responsibility. By exploring expansions (attributing partial agency to machines) and re-imaginings (positive responsibility, meaningful human control, accountability cultures), it demonstrates that **responsibility in computing** is neither purely an individual matter nor purely a machine matter—rather, it is a *hybrid phenomenon* emerging from the interplay between people, organizations, and technology.

David J. Gunkel (2020) paper titled “Mind the gap: responsible robotics and the problem of responsibility.”

---

1. Introduction: Responsibility as the Ability to “Answer For…”

Gunkel starts by noting that we’re in an era of “responsible robotics,” where the question of *who* or *what* is to be held responsible for robot actions has become urgent. He ties this to Paul Ricoeur’s remark that “responsibility” is messy, spanning both civil and penal liability, as well as moral accountability. Gunkel says the question “Who (or what) can or should answer for a robot’s actions?” is central to any robust notion of “responsible robotics.” □cite□turn2file0□

He clarifies that the paper aims to:

1. **Diagnose** how conventional notions of responsibility are used in contexts where robotic decision-making is increasingly visible.
2. **Consider** how certain developments in robotics/AI complicate that conventional approach.
3. **Evaluate** possible frameworks (instrumentalism, machine ethics, hybrid approaches) for dealing with these complexities.

The analysis is chiefly *critical* rather than *normative*. That is, Gunkel wants to highlight the conceptual difficulties rather than propose a single best solution.

---

2. The Default “Tool” Interpretation

Gunkel’s first main section explores how we typically treat technology as *mere tools or instruments* in moral philosophy. This perspective, which he calls **instrumentalism**, rests on two claims:

> 1. **Technology as means to ends**
> 2. **Technology as subordinate to human intentions** (i.e., “human activity designs, deploys, and uses technology”).

He cites Martin Heidegger (1977) and Andrew Feenberg (1991) to underscore that instrumentalism is “the most widely accepted view of technology” □cite□turn2file0□. In its simplest form, it says: **only humans have moral agency**; technology is the neutral set of tools or instruments we employ for our purposes.

In moral responsibility terms, this means that when something goes awry (e.g., a malfunctioning self-driving car, or a lethal decision by an “autonomous” system), the question reduces to: “Which human(s) designed, used, or otherwise controlled this technology?” *They* are the ones who must “answer for” (Ricoeur’s phrase) the outcome. Or, if no humans can be singled out, we end up with "nobody is responsible," though Gunkel acknowledges that is often unsatisfying.

Philosophical justifications for instrumentalism:

- *Logical consistency*: Tools are, by definition, inert objects without ends of their own. They cannot "own" moral accountability.
- *Moral hazard argument*: If we blame technology itself, then we let human operators and designers evade accountability. They might say "the computer did it!" or "the robot glitched!"—which leads to the "computer as scapegoat" problem (Nissenbaum 1996).

Thus, if we do not maintain the principle that all moral accountability must remain on human shoulders, we might end up with irresponsible deflections. The adage "a poor carpenter blames his tools" captures this line of thought.

---

3. The "Robot Apocalypse": When Instrumentalism Isn't Enough

Next, Gunkel outlines **three** kinds of recent (or emerging) robotic / AI phenomena that stress-test the instrumentalist paradigm:

3.1 Autonomous Technology

Borrowing from Langdon Winner (1977), Gunkel discusses **autonomous machines**—mechanisms intended to replace human operators, rather than be used as mere tools. For instance, a self-driving car is not a *new car* so much as a *new driver*. It's not that we replaced the existing tool (the car) with some better tool. Rather, we replaced the *human's role* of driving with a machine occupant of that role. Hence:

> "Calling it a 'tool' might be factually off-target. In some sense, it is a 'machine' that replaces the human operator."

Gunkel points to self-driving vehicles, where the U.S. National Highway Traffic Safety Administration (NHTSA) concluded that the Google Car's **"Self Driving System"** could, for regulatory purposes, be deemed the vehicle's *"driver."* Hence, from a legal perspective, the occupant is *not* the driver; the occupant is something else (like a passenger), whereas the AI system is conceptually the "driver." That challenges the assumption that technology is "just a tool" in the driver's hand.

3.2 Machine Learning

He then highlights **machine learning**—algorithms that produce actions "out of our hands," i.e. not explicitly programmed in each step. Examples:

- **AlphaGo**: The Go-playing system that *learned* from data and self-play, surprising even its creators with unorthodox moves.
- **Microsoft Tay**: The chatbot that learned from Twitter interactions and quickly produced racist tweets, to Microsoft's embarrassment.

In both scenarios, the system is designed to do things *its* programmers did not (and could not) fully specify. Google's engineers didn't *manually encode* the strategies that beat Lee Sedol. Microsoft's engineers certainly didn't *intend* racist utterances. This means the old question "Who's at fault?" becomes tangled. Although we can still trace ultimate oversight to human beings, Gunkel stresses that:

> "The system is intentionally built to exceed the creators' direct oversight. That's the whole point of 'learning' algorithms."

Such outcomes reveal a **"responsibility gap"** (term from Andreas Matthias, 2004). The old tool-based approach struggles because the system's behavior can't be neatly predicted or directed by a single programmer's intentions.

3.3 Social Robots

Finally, Gunkel discusses **social robots**, like Cynthia Breazeal's Jibo. The Jibo marketing frames it as neither a mere "thing" (like a fridge) nor a fully recognized social family member, but *somewhere in between*. Gunkel draws parallels to how we treat pets or how some soldiers treat a bomb-disposal robot: they name it, bond with it, even risk their own safety for it. This complicates purely instrumental relationships.

The CASA (Computers As Social Actors) research by Reeves & Nass (1996) confirms that humans spontaneously treat anthropomorphic or interactive systems as if they had social standing. That doesn't necessarily *prove* the machine is a moral agent. Instead, it reveals that humans *perceive* social robots differently from how they perceive standard tools. This creates confusion over whether it is correct to speak of them *only* as neutral instruments.

---

4. Three Ways to Fill the Responsibility Gap

Having shown how some robots (or AI systems) strain the old approach, Gunkel outlines three possible responses:

4.1 Instrumentalism 2.0 (Strictly Reaffirm the Tool Paradigm)

We can **double down** on the idea that all technology is "just a tool," though increasingly complex. Gunkel cites Joanna Bryson's argument, "Robots Should be Slaves" (2010), which contends that morally or legally, we should treat robots as property: designing them to be subservient instruments. In effect, this reaffirms:

> "Responsibility remains firmly on humans. Even advanced systems are made by us, so they remain our slaves."

Advantages:

- Maintains clarity in moral and legal accountability (no "robot scapegoats").
- Aligns with existing product-liability doctrines.

Disadvantages:

1. Could hamper innovation, because if humans face total liability for unanticipated outcomes, they might not deploy advanced AI.
2. Deems "slavery" to be the default relation to machines, which might be psychologically or socially disturbing when people anthropomorphize them (like a dog or an empathetic caretaker). People *develop feelings* for these machines. Some worry this might degrade how we treat each other.

Essentially, Gunkel says this approach can become “slavery 2.0,” a new class of sub-beings. It might be logically consistent but can create social or moral friction when the machines display quasi-human or empathic behaviors.

4.2 Machine Ethics (Attribute Quasi-Responsibility to Robots)

Another response is to say “maybe some advanced robots can be recognized as *moral agents* or quasi-agents.” On that basis, we can embed them with *moral constraints* (machine ethics) and hold them partly accountable—just as we hold corporations legally accountable though they are artificial entities.

Wallach & Allen (2009) propose that we must design “moral machines” to handle potentially catastrophic decisions. **Anderson & Anderson (2011)** argue that well-programmed AI might handle ethical complexities *better* than inconsistent humans. The main idea: we see it all the time in corporate law, where a corporation is recognized as a “legal person.” Similarly, an AI might be recognized as an artificial moral agent for convenience.

However, Gunkel warns about pitfalls:

- Such a “machine morality” might produce “artificial bureaucrats” or “artificial psychopaths” that simply follow rules with zero empathy, arguably a substandard form of moral reasoning.
- It forces a deep rethinking of “human exceptionalism.”

Hence, while machine ethics is a real approach, it has nontrivial conceptual and practical consequences—*including* the possibility that these rule-bound robots could end up following rules blindly, ignoring contextual or empathetic nuance that humans might find essential.

4.3 Hybrid Responsibility (Distribute Across Human + Machine Networks)

A third possibility is to **distribute responsibility** across a socio-technical network, an idea Gunkel associates with:

- Actor-Network Theory (Latour 2005).
- “Extended agency” or “joint responsibility” (Hanson 2009).
- Verbeek’s “ethics of things” and Deborah Johnson’s triad (designer–system–user).

In this approach, responsibility is not pinned on a single agent (whether a person or the robot) but recognized as *emergent from the interplay* of all components. This acknowledges real-world complexity—for instance, the “many hands” problem that arises in climate change or large engineering projects. Instead of fixating on “who’s at fault,” the network perspective tries to see how responsibilities are distributed.

Drawbacks:

- Might diffuse blame to the point that *no one* can be pinned with accountability (like the 2008 financial crisis).
- Still demands that some authority or moral system decide which part of the network “counts” as an accountable agent, vs. which are mere background constraints.

Nevertheless, Gunkel sees this as a popular, flexible route for matching the complexities of advanced technology.

---

5. Concluding Observations: Why the Decision Matters

Gunkel finishes by emphasizing that **none** of these three responses is trivially correct or wrong. Each has trade-offs. The bigger point is that society must choose carefully **how** we conceptualize advanced robotics/AI in order to maintain a coherent approach to responsibility:

> “We are responsible for deciding who or what is a moral subject, and we are responsible for the consequences of that decision.”

Thus, “responsible robotics” is not only about ensuring safety or reliability; it’s also about how we draw moral lines in a world where technology is no longer neatly subordinate to each user’s direct commands. The “gap” introduced by autonomy, learning, and social presence forces us to reevaluate both philosophical and legal frameworks for accountability. And, crucially, how we close that gap affects not just robots, but our entire social fabric.

Key Takeaways

1. **Instrumental Theory**: Historically robust, but now challenged by autonomous behavior, learning algorithms, and anthropomorphized social robots.
2. **Responsibility Gap**: Emerging from the partial unpredictability and user’s emotional engagement with AI systems.
3. **Three “Solution Clusters”**:
   - Reassert the tool stance (robots as property/slaves).
   - Embrace machine ethics (robots as new moral/quasi-agents).
   - Hybrid distribution of responsibility in socio-technical networks.

By presenting these divergences clearly, Gunkel’s paper fosters deeper discussion of what “responsible robotics” might look like in practice—and how each approach reshapes the moral landscape.