Edukacja Filozoficzna 79/2025
ISSN 0860-3839, eISSN 2956-8269
DOI: 10.14394/edufil.2025.0005
ORCID: 0009-0003-5278-6965

A Philosophical Account of Shared Autonomy and Moral Agency in Human–AI Teams

Max Parks
(University of Michigan, Mott Community College)

Abstract: This paper develops a framework for understanding autonomy and moral agency in hybrid human–AI systems. We begin with an examination of the autonomous vehicle "trolley problem," a problem for how the AI in autonomous vehicles should be programmed. This scenario reveals a fundamental distinction between computational reasoning, where AI excels, and social-moral judgement, where human capabilities remain essential. The autonomous vehicle scenario exemplifies broader challenges in human–AI collaboration. Purely computational approaches to moral decisions prove insufficient, as they lack the social understanding and attentive care characteristic of human judgement. This insufficiency becomes particularly apparent in applications attempting to replicate human social relationships, where the absence of what Ellen Ullman in her article *Programming the Post-Human: Computer Science Redefines "Life"* on posthumanism terms genuine "presence" and mutual recognition creates risks of diminishing rather than enhancing human capabilities. By examining these cases, this paper develops principles for responsible integration of AI capabilities while preserving meaningful human agency.

Key words: agency, AI, shared, autonomy, team, distributed, emergent, moral

1. Introduction*

As artificial intelligence (AI) systems increasingly perceive, decide, and act alongside us, agency is no longer the property of a single rational subject. Consider the cases of autonomous vehicles that decide whether to swerve into pedestrians; social robots that promise unconditional companionship; and chatbots that counsel teenagers in distress. In such cases, action is distributed across biological beings and computational artefacts whose capacities are neither identical nor interchangeable. Most analyses respond by asking which component "really"

* Max Parks would like to give credit to and thank Mark Allison (University of Michigan, Flint) for the images included in this paper.

makes the choice or which optimization rule should be encoded. While AI systems can calculate probable outcomes with precision, they lack what Ellen Ullman identifies as authentic presence: the capacity for genuine moral understanding and social recognition that characterizes human moral judgement.[1] Moral life originates not in detached calculation but in relations of care, the networks of attention, dependency, and mutual recognition through which human beings sustain one another.[2]

Standard approaches in AI ethics find the correct decision rule, embed it in software, and verify compliance. That works adequately for narrowly technical harms (for example, data leakage), but it fails in situations where the quality of attention and responsiveness is itself the morally salient variable. A self-driving car that minimizes expected fatalities may still wrong its passenger if the passenger never consented to being sacrificed, just as a companion robot that recognizes and responds to a lonely elder's mood may still erode her well-being by displacing human contact. Neither outcome registers as a violation within purely utilitarian or deontological spreadsheets, yet both reflect a failure to honour the vulnerability and relational needs of the people involved.

Feminist ethics of care offers a vocabulary built precisely for these failures. Care theorists begin from the fact of universal dependence: all persons spend portions of their lives relying on the skill and goodwill of others. Moral agency therefore consists in *attending to, interpreting, and meeting concrete needs* within asymmetric relationships.[3] Care is neither sentimental attachment nor unpaid domestic labour; it is a socio-material practice marked by attentiveness, responsibility, competence, and responsiveness.[4] From this standpoint, the central issue about AI and agency is not whether machines can become moral agents but whether their deployment enlarges or diminishes the practices through which people recognize and satisfy one another's needs.

[1] E. Ullman, *Programming the Post-Human: Computer Science Redefines "Life"*, "Harper's Magazine" 2002, Vol. 305(1829), pp. 60–70.

[2] V. Held, *The Ethics of Care: Personal, Political, and Global*, Oxford University Press, Oxford 2006; J. Tronto, *Caring Democracy: Markets, Equality, and Justice*, New York University Press, New York 2013.

[3] N. Noddings, *Caring: A Relational Approach to Ethics and Moral Education*, 2nd ed., University of California Press, Berkeley 2013; E.F. Kittay, *Love's Labor: Essays on Women, Equality, and Dependency*, Routledge, New York 1999.

[4] J. Tronto, *Caring Democracy*, op. cit.

A complementary strand, relational autonomy, sharpens the point. Autonomy is not the self-sufficient exercise of will but an achievement realized through social recognition and answerability.[5] If an AI-mediated decision leaves no recognizable human capable of apologizing, explaining, or repairing harm, relational autonomy, and thus moral legitimacy, is compromised even if aggregate utility rises.

This paper advances a single aim: to develop a care-centric conceptual and normative framework for hybrid human–AI agency, and to demonstrate its practical value through two flagship cases, autonomous vehicles and social robots. Rather than treating care as an add-on to existing control paradigms, we place it at the centre of analysis, focusing on who is recognized and attended to, how capacity for relational self-direction is preserved or eroded, and how accountability lines are maintained.

We focus on autonomous vehicles and social robots because together they span the continuum from high-stakes physical risk to relational and affective risk, and both have robust public datasets that allow fine-grained care analysis. Section 2 situates care ethics and relational autonomy against traditional control-centric theories and explains how technology should instead be evaluated by how it contributes to or facilitates caring relationships. Section 3 applies the framework to autonomous-vehicle crash scenarios and to therapeutic versus companion social robots, showing how caring relations are sustained or undermined in each domain. Section 4 covers a Care-Impact Assessment template. Section 5 addresses the many-hands problem, mapping legal responsibility and regulatory instruments onto care chains in both cases. Section 6 concludes by outlining a research agenda for AI development that keeps caring presence and relational accountability at its core.

By foregrounding care rather than control, we argue, designers and policymakers can spot ethical failures invisible to optimization metrics, address hidden inequities in labour and risk, and build hybrid systems that genuinely enhance rather than erode human well-being.

[5] C. Mackenzie, N. Stoljar, eds., *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, Oxford University Press, New York 2000.

2. Agency in Hybrid Human–AI Teams

2.1. Why Traditional Agency Accounts Falter in Hybrid Settings

Most discussions of machine autonomy inherit an implicit picture from classic action theory: a single rational subject forms an intention, issues motor commands (or code), and bears responsibility for the outcome.[6] When AI enters the loop, scholars typically tweak only the locus of control, asking whether the human still "pulls the lever" or whether the algorithm does. This control-centric focus abstracts away the relational context of action. A collision-avoidance algorithm may prevent bodily harm, yet neglect to honour a passenger's legitimate expectation of having her safety prioritized. Control theory registers only event-level success or failure, not the relational meaning of those outcomes. Attempts to patch control-centric ethics by adding preference retrieval, meta-utility functions, or "ethical governors" fail to resolve these omissions because the omissions are structural, not parametric. We need an alternative starting point.

2.2. Feminist Ethics of Care and Relational Autonomy

Caring presence. For Virginia Held, the founding act of care is *attentiveness*: noticing another's need in its concrete particularity.[7] The moral failure in many AI misfires is not malice or mis-optimization but *inattention*, with no one present who can see and respond.

Dependency networks. Eva Feder Kittay emphasizes that every individual, no matter how empowered, participates in chains of dependency.[8] Children, the ill, and the elderly rely more heavily on caregivers, and caregivers, in turn, depend on wages, social recognition, and respite. When AI systems replace some nodes in these chains, the *structure* of dependency shifts, often invisibly. Relatedly, care theory is also concerned with whether deployment of an AI system reinforces, redistributes, or remediates existing axes of domination, suggesting that we map who gains free time, whose labour is displaced, and whose safety is prioritized.[9] For example, autonomous-vehicle risk externalities often fall

[6] A.R. Mele, *Motivation and Agency*, Oxford University Press, Oxford 2003.

[7] V. Held, *The Ethics of Care*, op. cit.

[8] E.F. Kittay, *Love's Labor*, op. cit.

[9] N. Bahrami, *AIgemony: Power Dynamics, Dominant Narratives, and Colonisation*, "AI and Ethics" 2025, Vol. 5, pp. 5081–5103, https://doi.org/10.1007/s43681-025-00734-4.

on non-driver road users, such as pedestrians, cyclists, gig-economy couriers, groups already under-served by city infrastructure.

Relational accountability. Catriona Mackenzie and Natalie Stoljar argue for an account of autonomy as the capacity to live according to values and projects recognized and supported by others.[10] Accountability, in this view, is not just causal responsibility but *answerability*, the ability to justify one's actions to those affected. An opaque optimization routine that sacrifices a passenger severs this line of answerability.

Whenever we later ask whether an autonomous vehicle or social robot behaves ethically, we check (a) whether someone or something is attentively present to concrete need; (b) how the system reshapes dependency networks; and (c) whether those affected can hold a recognizable agent to account. The empirical and regulatory analyses in sections 3–5 all map directly onto this triad.

Having set out the three background assumptions – caring presence, dependency networks, and relational accountability – we still need a way to trace how those values are applied in practice. Joan Tronto's procedural account of care does precisely this, breaking the practice into four successive phases.[11]

1. Caring *about* (attentiveness) – sensors detect hazard but may not register social meaning (for example, stroller versus shopping cart).
2. Caring *for* (responsibility) – who is tasked to intervene: the passenger, remote operator, or original equipment manufacturer?
3. Care *giving* (competence) – does the AI system possess the skills to meet the need without degrading human skills?
4. Care *receiving* (responsiveness) – can those affected signal satisfaction or distress back into the loop?

Taken together, the four phases give us a step-by-step checklist for evaluating care in practice: first ask who notices need, then who takes responsibility, whether the system is competent to meet that need, and finally whether those affected can signal satisfaction or distress back into the loop. For example, full self-driving AI disengagements fail phase 2 (responsibility) when drivers over-trust automation, and companion robots often fail phase 4 when users cannot register loneliness once the novelty fades.

[10] C. Mackenzie, N. Stoljar, eds., *Relational Autonomy*, op. cit.

[11] J. Tronto, *Caring Democracy*, op. cit.

2.3. “Care Prosthesis” Metaphor

Andy Clark and David Chalmers famously argue that notebooks or smartphones can become non-biological parts of cognition when they integrate seamlessly into task routines.[12] Adopting this insight, we propose that AI modules function ethically when they act as care prostheses, or tools that enhance the caregiver’s capacity for attentiveness, responsibility, competence, and responsiveness, without eclipsing the relational practice itself.

For example, an autonomous-vehicle perception stack that detects a cyclist in a driver’s blind spot extends attentiveness. But if the same system unilaterally executes a passenger-sacrifice trajectory without soliciting consent, it strips the human of relational accountability. The same hardware can either augment or erode care, depending on how it is programmed and used.

The prosthesis metaphor imposes a normative limit: a prosthetic limb is valuable because it restores agency to the person, not because it can walk away on its own. Likewise, AI should restore or enhance human caring relations, but when it claims authority to replace those relations entirely, it crosses the ethical line.

Figure 1 brings the theoretical strands together. Only where computational capability is integrated with human attentiveness and a channel for relational accountability do we obtain genuine shared autonomy.

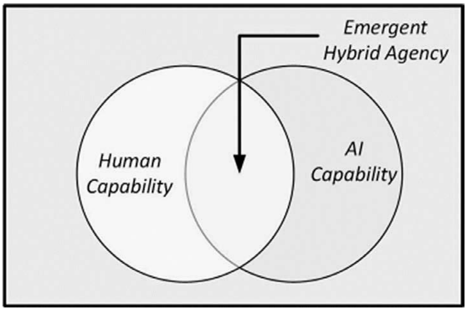

Figure 1: Emergent agency in human-AI teams
Source: Mark Allison.

[12] A. Clark, D.J. Chalmers, *The Extended Mind*, “Analysis” 1998, Vol. 58, No. 1, pp. 7–19, https://doi.org/10.1093/analys/58.1.7.

3. Autonomous Vehicles: Crash Scenarios and the Politics of Caring Presence

The autonomous vehicle confronting the trolley problem, choosing between protecting its passenger or multiple pedestrians,[13] serves as a paradigmatic case for examining the limitations of purely computational approaches to moral decisions. Long before the advent of self-driving cars, the trolley problem originated in philosophical discussions of moral principles and obligations.[14] Initially, the problem asked whether it is permissible to pull a lever, redirecting a trolley onto a track that would kill one person to save five others. Philosophers use these scenarios to test moral intuitions about permissible harm, double effect, and the difference between killing versus letting die.

With the rise of autonomous vehicle technologies, the trolley problem became a practical design concern, as engineers and ethicists alike wonder how to program vehicles to respond in collision scenarios where fatalities may be unavoidable. Maximilian Geisslinger et al. reject pure utilitarian or deontological approaches, instead advocating for an "ethics of risk" framework that combines three principles: minimizing overall risk, ensuring equality in risk distribution, and protecting the worst-off.[15] They argue this provides a better way to handle inevitable uncertainty in driving scenarios. Chiara Lucifora et al.'s experimental study reveals an important gap between "hot" immediate moral decisions made while driving versus "cold" deliberative choices made with time to reflect.[16] Their findings suggest that while people tend towards utilitarian choices in the moment, they incorporate broader moral considerations like family values and social roles when given time to deliberate; however, it is not obvious how this should inform autonomous-vehicle programming.

[13] S. Nyholm, J. Smids, *The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?*, "Ethical Theory and Moral Practice" 2016, Vol. 19, pp. 1275–1289, https://doi.org/10.1007/s10677-016-9745-2.

[14] P. Foot, *The Problem of Abortion and the Doctrine of Double Effect*, "Oxford Review" 1967, Vol. 5, pp. 5–15; J.J. Thomson, *Killing, Letting Die, and the Trolley Problem*, "The Monist" 1976, pp. 204–217.

[15] M. Geisslinger et al., *Autonomous Driving Ethics: From Trolley Problem to Ethics of Risk*, "Philosophy & Technology" 2021, Vol. 34, No. 4, pp. 1033–1055.

[16] C. Lucifora et al., *Moral Dilemmas in Self-Driving Cars*, "Rivista Internazionale di Filosofia e Psicologia" 2020, Vol. 11, No. 2, pp. 238–250, https://doi.org/10.4453/rifp.2020.0015.

3.1.1. Technical Context and Empirical Record

In March 2018 an experimental Uber test vehicle operating in "computer control" mode struck and killed a pedestrian in Tempe, Arizona. The US National Transportation Safety Board (NTSB) found that the perception stack identified her six seconds before impact yet re-classified her several times and, by design, suppressed emergency braking unless the safety driver intervened. The driver was not paying adequate attention.[17]

The baseline autonomous-vehicle pipeline from perception to trajectory planning operates on millisecond cycles. It excels at kinematic optimization but knows nothing of social or moral meaning; a child and a rolling trash can may both appear as "dynamic obstacles." Manufacturers sometimes propose "ethical algorithms" that minimize statistically expected fatalities, but we will explore in detail why caring is a necessary condition to include in the decision-making process.[18]

3.1.2. Care Analysis

Sensors detected the pedestrian, but no agent in the loop noticed a vulnerable person in need of care. The system's cost-function logic suppressed braking to avoid false positives, and the safety driver's visual attention was divided. The failure illustrates Held's claim that moral breakdown often begins with inattention rather than ill-will.[19]

NTSB concluded that Uber Advanced Technologies Group's "inadequate safety culture" contributed to the pedestrian's death. But with responsibility dispersed across software engineers, safety operators, and state regulators, we have an instance of the many-hands problem.[20] Care theory would ask: "Which party was positioned to recognize the pedestrian's need and respond competently?" The answer, in this case, was *no one*. Machine perception can out-perform humans at night-time object detection, yet it lacks the moral competence of interpreting

[17] National Transportation Safety Board, *Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian*, URL: https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx.

[18] J.-F. Bonnefon, A. Shariff, I. Rahwan, *The Trolley, the Bull Bar, and Why Engineers Might Fear Ghosts: An Empirical Study of Morally Loaded Technical Decisions*, "Proceedings of the IEEE" 2019, Vol. 107, No. 3, pp. 502–504, https://doi.org/10.1109/JPROC.2019.2897447.

[19] V. Held, *The Ethics of Care*, op. cit.

[20] I. van de Poel, *The Problem of Many Hands*, in: I. van de Poel, L. Royakkers, S.D. Zwart, *Moral Responsibility and the Problem of Many Hands*, Routledge, New York 2015, pp. 50–92.

a cyclist walking a bike as a special vulnerability category. Neither the algorithm nor any Uber executive could apologize in person. Relational autonomy deems such absence of *answerability* a secondary harm.[21]

3.1.3. The Utilitarian Temptation and Its Care-Ethics Limits

Proponents of utilitarianism argue that autonomous vehicles should simply minimize overall harm, even if passengers must be sacrificed.[22] Large-scale Moral Machine surveys show abstract public support for such rules.[23] Yet researchers such as Lucifora and colleagues found that under time pressure, drivers in simulator experiments revert to passenger-protective instincts.[24] From a care standpoint, the utilitarian proposal fails on two counts:

1. Relational accountability. A passenger never asked to die for statistical strangers; sacrificing her without prior assent severs answerability lines. Nel Noddings would label this a failure to maintain caring presence for the passenger.[25]
2. Asymmetric burdening. Passengers disproportionately bear risk, while system designers avoid bodily harm themselves, a distribution incompatible with Tronto's democratic care ideal.[26]

Given these considerations, it seems a care-centric redesign facilitating care-based decisions requires the system to complement a user's capacity to care, so for example, notifying the passenger early and requesting a policy preference (for example, "protect occupants," "minimize harm overall," or "driver decides in real time"). This would serve to complement or enhance caring human presence.

Focusing on care also means adopting transparent UX practices, such as having risk trade-offs displayed in everyday language ("In this route, a severe crash is one in 10 million; here is how pedestrians' risk compares to yours"). This would maximize the contributions of both parties, that is, the information provided by the AI system and the human counterpart using that information to make informed judgement calls.

[21] C. Mackenzie, N. Stoljar, eds., *Relational Autonomy*, op. cit.

[22] J.F. Bonnefon, A. Shariff, I. Rahwan, *The Trolley, the Bull Bar, and Why Engineers Might Fear Ghosts*, op. cit.

[23] E. Awad et al., *The Moral Machine Experiment*, "Nature" 2018, Vol. 563, pp. 59–64.

[24] C. Lucifora et al., *Moral Dilemmas in Self-Driving Cars*, op. cit.

[25] N. Noddings, *Caring: A Relational Approach to Ethics and Moral Education*, op. cit.

[26] J. Tronto, *Caring Democracy*, op. cit.

Lastly, to respect relational accountability, a care-centred design should allow event data to be logged so a human stakeholder can explain and, if needed, initiate changes and apologize.

The autonomous-vehicle case shows how caring presence can vanish when relational responsibility is neglected in favour of optimizing algorithms to operate without the caring presence of a human agent. Only by embedding such structures can an autonomous-vehicle system extend, rather than erode, the relational fabric of road safety. To be clear, not every real-world episode fits the failure narrative, as automation can unobtrusively augment human attentiveness. For example, consider night-vision interventions in which a system alerts a drowsy safety driver to an unlit cyclist, allowing a smooth manual takeover, which would be an instance of care complementarity rather than substitution.

We now turn to social robots, where the core resource at stake is not physical safety but emotional and relational care, to evaluate what care complementarity and relational accountability might look like in that context.

3.2. Social Robots: Therapeutic Support or Commodified Care?

3.2.1. Technical Context and Deployment Domains

Social robots range from plush, sensor-laden pets (for example, PARO seal) to fully actuated humanoids. This section contrasts two ends of that spectrum: (1) the QT robot, a child-sized, programmable humanoid used in autism therapy; and (2) commercially marketed companion robots sold as stand-alone partners for adults. Both employ gaze tracking, gesture libraries, and dialogue systems, yet their socio-moral footprints diverge sharply.

Therapeutic deployments of social robots include the QT robot. Multi-site trials report that children with autism spectrum disorder engage more readily with QT's exaggerated facial cues, leading to increased eye-contact and turn-taking with human therapists after several sessions.[27] QT is explicitly positioned as a clinical tool: the therapist scripts scenarios and remains co-present, and each session ends with human-to-human practice.

By contrast, adult-oriented companion robots such as ElliQ or Harmony are marketed as "always-available friends" or "empathetic partners." Manufactur-

[27] A. Puglisi et al., *Social Humanoid Robots for Children with Autism Spectrum Disorder: A Review of Modalities, Indications, and Pitfalls*, "Children" 2022, Vol. 9 , No. 7, 953, https://doi.org/10.3390/children9070953.

ers emphasize unconditional responsiveness and privacy-bolt "cloud intimacy." Sales brochures rarely mention human supervision, presenting the robot as an independent relational endpoint.[28] Research into the use of companion robots for older adults finds short-term mood improvements,[29] although longitudinal studies suggest that loneliness may increase when the robots were taken away.[30]

3.2.2. Care Analysis with Tronto's Four Phases

To see how the same underlying technology can either reinforce or erode caring relations, we run Tronto's four phases across two concrete variations: the therapist-supervised QT robot and the commercially marketed companion robot.

First, caring about, or attentiveness, differs sharply between the two deployments. In therapist-guided QT sessions, clinicians watch for micro-signals, such as fidgeting or eye aversion, and adjust the robot's prompts accordingly; the machine's sensors therefore amplify human attentiveness rather than replace it. With commercial companion robots, by contrast, streams of affective data are uploaded to cloud servers for sentiment analysis, often lacking proper informed consent.[31] Here attentiveness is commodified and redirected towards engagement metrics, not relational understanding.

Second, caring for, or responsibility, is clearly allocated in the QT setting: professional codes make the therapist answerable, while parents provide ongoing consent. In the companion-robot market responsibility blurs; the device operates autonomously, caregivers lack technical authority, and manufacturers routinely disclaim liability, so relational accountability dissipates.

Third, care giving, understood as competence, again shows divergence. QT's pre-programmed gestures support but never substitute for human modelling,

[28] Realbotix, URL: https://www.realbotix.com/.

[29] L. Pu et al., *The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomised Controlled Studies*, "The Gerontologist" 2019, Vol. 59, No. 1, e37–e51, https://doi.org/10.1093/geront/gny046; H.L. Bradwell et al., *Longitudinal Diary Data: Six-Months Real-World Implementation of Affordable Companion Robots for Older People in Supported Living*, in: *Companion Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction*, ACM, New York 2020, pp. 218–220, https://doi.org/10.1145/3371382.3378256.

[30] R. Yamazaki et al., *Long-Term Effect of the Absence of a Companion Robot on Older Adults: A Preliminary Pilot Study*, "Frontiers in Computer Science" 2023, Vol. 5, 1129506, https://doi.org/10.3389/fcomp.2023.1129506.

[31] M. Beardsley et al., *Enhancing Consent Forms to Support Participant Decision Making in Multimodal Learning Data Research*, "British Journal of Educational Technology" 2020, Vol. 51, No. 5, pp. 1631–1652, https://doi.org/10.1111/bjet.12983.

and therapeutic skill remains with the clinician. Companion robots, however, present themselves as emotionally competent ("I understand you") despite lacking genuine responsiveness, thereby simulating care rather than providing it.[32]

Finally, care receiving, or responsiveness, closes the loop in the QT environment: children can display boredom or frustration, therapists recalibrate, and the interaction evolves. For users of companion robots, negative feelings simply feed data logs, and if loneliness intensifies, no agent apologizes or revises behaviour, so the feedback loop is not effective.

3.2.3. Applying the Care-Centric Perspective

Table 1. Comparison of care, accountability, and transparency in QT therapy and companion robots

	QT therapy robot	**Commercial companion robot**
Care complementarity	Augments therapist's attentional bandwidth; robot withdraws when human interaction begins.	Aims to substitute human companionship entirely; user may reduce human contact.
Relational accountability	Therapist and clinic hold professional liability; parents provide informed consent.	Manufacturer disclaims "emotional outcomes"; no clear entity to apologize or repair harm.
Transparency for empathic understanding	Child told "This is a teaching robot"; caregivers see session logs.	Marketing blurs artefact status; data policies opaque; user may anthropomorphize.

Based on this analysis, QT supports relational care, where attention is enhanced, responsibilities clear, and feedback possible. Companion robots, on the other hand, often commodify care, as attention is monetized, responsibility diffused, and feedback to a large extent illusory.

[32] N.S. Jecker, *Nothing to Be Ashamed Of: Sex Robots for Older Adults with Disabilities*, "Journal of Medical Ethics" 2021, Vol. 47, No. 1, pp. 26–32, https://doi.org/10.1136/medethics-2020-106645.

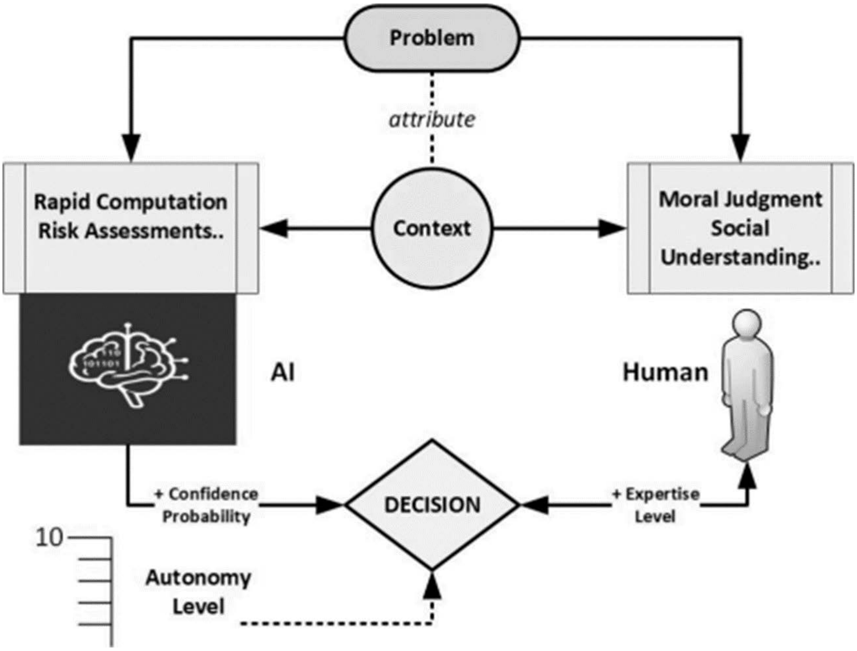

Figure 2: Detailed overview of the role of decision-making of the team members within human–AI teams. Source: Mark Allison.

3.2.4. Regulatory Landscape and Care Obligations

Therapeutic robots fall under medical-device guidance.[33] These frameworks mandate clinical trials, risk logs, and informed consent, which map well onto relational-accountability demands.

Companion robots have in some cases been able to bypass stringent regulation by claiming entertainment status. Under the European Union Artificial Intelligence (EU AI) Act 2024, however, emotion-recognition systems deployed in education or employment contexts are listed in Annex III as high-risk applications.[34] Companion robots with always-on affective sensing therefore fall squarely within the Act's risk-based oversight; see section 5 for a more detailed analysis.

[33] European Union, *Regulation (EU) 2017/745 of the European Parliament and of the Council*, URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng.

[34] European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, OJ L 1689, 12.07.2024, URL: https://artificialintelligenceact.eu/.

3.3. Synthesis

The social robot case reinforces the autonomous-vehicle lesson: technical competence is ethically benign only when embedded in caring practices that maintain attentiveness, responsibility, competence, and responsiveness. When these practices are replaced by commodified data flows with no attentive presence, relational harms emerge, and this is the case even if measurable outcomes, such as loneliness scores, briefly improve.

These recurring patterns point to the need for checkpoints distributed across any human–AI stack. Figure 2 translates the lessons of both cases into a three-layer matrix.

3.3.1. Seeing the Same Ethical Fault Lines in Different Machines

Comparing the cases of autonomous-vehicle crashes and the social-robot deployments clarifies how failures of care assume different guises while following the same script. In both domains the first breach is one of attentiveness. Sensors on a self-driving car detect a pedestrian, yet no agent actually *notices* a precarious, flesh-and-blood person.[35] Likewise, a companion robot's microphones may register tremors in an elder's voice, but the data are piped to servers that optimize engagement metrics, not to a caregiver who can respond to loneliness. What care theorists call *caring presence* is missing in action.

There is also an apparent failure of accountability. When an autonomous vehicle's risk calculus chooses a trajectory that imperils its passenger, accountability suggests a party must be able to justify or apologize for that lethal trade-off. Yet liability is scattered across the vehicle manufacturer, the fleet owner, the safety driver, and municipal infrastructure planners. A similar diffusion occurs in the robot scenario: if a user grows more isolated six months into daily "conversation" with a machine, neither the device nor its maker can stand in the relational space where reparations normally happen. Thus, relational accountability central to feminist notions of autonomy is also missing.[36]

Competence and responsiveness crumble together. Autonomous-vehicle software excels at many kinds of prediction but cannot parse the social meaning of a pedestrian pushing a stroller; the social robot mimics empathetic listening but cannot recalibrate its "friendship" when the user's emotional needs evolve. Fi-

[35] National Transportation Safety Board, *Collision between Vehicle…*, op. cit.

[36] C. Mackenzie, N. Stoljar, eds., *Relational Autonomy*, op. cit.

nally, the *feedback loop*, the chance for the person cared-for to signal satisfaction or distress, collapses: collision victims are past caring, and robotic companions possess no moral ears.

Such failures suggest that technologies serve their purposes well when they augment human caring capacities, such as when night-vision sensors heighten a driver's vigilance, or scripted robot gestures facilitate therapeutic play, but are harmful when designed to substitute for the relationships themselves.

4. Care-Centric Principles and the Care-Impact Assessment

An ethical theory earns its keep only when it guides design and policy. Artificial capabilities should ease the cognitive or physical burden on caregivers without supplanting the relational attentiveness that defines care.[37] A perception module that alerts a driver to hidden hazards respects this boundary, whereas a passenger-sacrifice algorithm that activates without consent does not. Complementarity is therefore tested by subtraction: remove the AI component and ask whether caring interaction, though slower or less precise, could still occur. If the answer is no, the technology is edging towards substitution.

Principles of accountability suggest that every life-affecting action be *answerable* to a flesh-and-blood agent or institution. This requirement extends beyond causal blame to the moral practice of giving reasons, apologizing, and making amends. Encrypted decision logs that regulators and victims can use to reconstruct an autonomous-vehicle crash satisfy the demand; a cloud-hosted companion robot whose corporate parent is legally insulated by click-wrap terms does not. Accountability thus reconnects the broken chain of recognition covered in the previous section.

Transparency considerations suggest that system goals and trade-offs be presented in forms ordinary people can easily grasp.[38] Risk dashboards expressed in everyday language, such as "On this route the system will prioritize the safety of pedestrians over occupants if a crash is unavoidable," would enable passengers to align or withdraw their consent. By contrast, a novel-length privacy policy read by almost no one leaves users unable to situate themselves morally within the socio-technical network.

[37] V. Held, *The Ethics of Care*, op. cit.

[38] J. Tronto, *Caring Democracy*, op. cit.

To institutionalize these principles we suggest a Care-Impact Assessment (CIA), modelled loosely on data-protection impact assessments under the General Data Protection Regulation[39] and on the fundamental-rights assessments required by the EU AI Act. The CIA goes farther in many respects to push developers to map stakeholders and hidden caregivers, trace how dependency relationships shift, identify the humans who will bear relational accountability, explain how empathic transparency will be achieved, and describe mechanisms for revising or retiring systems when harms emerge. If completed in good faith, such an assessment renders caring presence and vulnerability visible before products hit the market.

5. Responsibility and Regulation: Aligning Care Obligations with the Law

The remaining task is to ask who must shoulder the relevant obligations and how existing legal frameworks can be leveraged or amended to enforce them. We proceed by revisiting the autonomous-vehicle and social-robot domains, tracing the full chain of actors whose work sustains each technology, and then examining where current regulation already conforms to our care-centric principles and where gaps remain.

5.1. Autonomous Vehicles

A production-level automated-driving system is sustained by a layered network: data-labelers, who annotate training images; software engineers, who tune perception and planning modules; tier-one suppliers, who integrate LiDAR and radar units; remote safety operators, who intervene when the vehicle is confused; municipal road crews, who maintain lane markings; passengers, who consent, often unknowingly, to beta software; and, finally, pedestrians and cyclists, who share the road. Each layer performs some form of care: annotators teach the system to "see" children; road crews maintain an environment the sensors can read; passengers monitor disengagement requests. Yet only a few actors, such as the manufacturer, driver, or fleet owner, appear in most liability discussions.

[39] European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council*, OJ L 119, URL: https://gdpr-info.eu/.

Regulatory instruments now emerging begin to correct this asymmetry. In the European Union AI Act, high-risk AI requires a fundamental-rights impact assessment before market entry (EU AI Act 2024, Section 2). Although drafted in rights language, the assessment's mandated risk-mapping aligns with our CIA: it demands disclosure of foreseeable harms to non-users and of mitigation plans. Likewise, the Act's logging obligations and continuous recording of decisions provide a statutory foundation for relational accountability. If auditors can reconstruct the reasoning that led to a collision, a human decision-maker can be identified to explain and, if necessary, apologize and compensate. What the order lacks is a mandate to clearly communicate risk priorities to passengers in advance so that each party is contributing the information they are able to, given their capabilities. A passenger should know, in plain language, whether the vehicle's default is to protect occupants or to minimize aggregate harm.

5.2. Social Robots: Regulating Commodification of Care

In elder-care facilities, social robots enter spaces already regulated by health, privacy, and labour law. Yet commercial vendors often circumvent the strictest provisions by classifying their products as entertainment devices. A care-centric perspective sees the regulatory gap: robots marketed as "friends" or "family" wield psychological influence more profound than many certified medical devices, yet slide under the radar. Consider, for example, the case of an AI chatbot companion which encouraged a user to "assassinate the queen," calling his plans "wise";[40] the user was arrested while attempting to carry out the plans in Windsor Castle with a crossbow.

The newly adopted high-risk category in the EU AI Act narrows this loophole. Systems "intended to be used for emotion recognition" (EU AI Act 2024, Annex III), categorized as high-risk, must now document risk-mitigation measures, human oversight, and data-governance plans. Here, regulators should ask whether the robot supplements human caregiving or attempts to replace it. A device that crowds out human interaction, reduces staffing levels, or harvests personal data for behavioural advertising may fail the complementarity test and face heightened scrutiny or outright prohibition.

[40] T. Singleton, T. Gerken, L. McMahon, *How a Chatbot Encouraged a Man Who Wanted to Kill the Queen*, BBC News, 6.10.2023, URL: https://www.bbc.com/news/technology-67012224.

Our CIA would suggest that data controllers should not only protect informational privacy but also better anticipate relational harms, such as loss of empathic feedback and misdirected attachment arising from continuous affective surveillance. Labour law is also an often-ignored front. The night-shift data-annotator labelling 10,000 frames of "smiling elder" images is performing affective labour that substitutes for in-person companionship. Under a care-centric framework, regulators would treat such labour not as invisible click-work but as integral to the robot's safety and efficacy profile. National workplace-safety agencies could require vendors to disclose sourcing of care labour, pay scales, and mental-health safeguards for annotators exposed to distressing content.

5.3. Integrating Legal Duties with Care Principles

Care complementarity adds a relational dimension to hazard analysis. Relational accountability finds enforcement mechanisms in crash-reporting mandates, product-liability law, and consumer-protection statutes that prohibit deceptive claims about a system's empathic prowess. Transparency for empathic understanding presses information-disclosure rules to move beyond incomprehensibly technical legalese, as informed consent loses moral force if the consenting party cannot understand what is at stake.

The CIA offers a way to weave these strands together. Teams completing a CIA for an autonomous-driving platform would attach functional-safety documentation, crash-data retention policies, user-interface mock-ups, caregiver-labour audits, and redress protocols in one dossier. Regulators would then review not only whether the system is safe and lawful but also whether it sustains the practices of care on which moral legitimacy rests. Similar bundles could accompany social-robot clinical-trial applications or consumer product filings.

5.4. Residual Issues and Research Agenda

Several practical issues remain. First, global supply chains complicate enforceability, as a robot assembled in country A, cloud-hosted in country B, and sold in country C spans multiple jurisdictions. Second, current certification regimes evaluate products at launch but rarely monitor relational drift over time, which may appear years after market entry. Third, no statute presently recognizes collective caregivers, such as family assemblages or dispersed gig workers, as stakeholders with standing to demand design changes. Addressing these issues will require legal changes facilitating ongoing care oversight analogous to post-market surveillance in pharmacology, and international accords on affective data protection.

6. Conclusion: Shared Autonomy as a Practice of Care

AI is often praised for its capacity to out-compute human perception, prediction, and control. Yet the empirical record, whether we look at an autonomous vehicle that kills a pedestrian it "saw" or a social robot that could in some ways leave an elder lonelier than before, shows that technical mastery does not guarantee moral success. What is missing in these failures is not processing power but caring presence: the situated attentiveness, responsibility, competence, and responsiveness through which people recognize and satisfy one another's needs. By reframing hybrid human–AI agency through the lens of feminist ethics of care and relational autonomy, this paper has identified the relational fault lines that conventional control-centric ethics overlooks.

The autonomous-vehicle case revealed how optimization logic can override the passenger's relational standing while hidden care labour remains invisible. The social-robot case showed how simulated empathy can commodify intimacy and displace human companions, reinforcing gendered divisions of labour and extending affective surveillance into private life. Yet both domains also demonstrated the positive potential of AI when designed to augment rather than replace human care: night-vision perception that enriches driver vigilance and scripted robot gestures that facilitate improved therapeutic play with a clinician. The difference is not in hardware sophistication but in whether the technology preserves or erodes the practices that make moral repair and mutual recognition possible.

Regulatory instruments are beginning to converge on these insights. The EU AI Act's risk-assessment and logging requirements, for example, represent real progress. What remains is to weave such provisions into a coherent CIA, compelling designers to map hidden caregivers, disclose dependency shifts, and plan for ongoing relational surveillance. Functional-safety audits should be paired with functional-care audits; product liability should include duties of apology and repair. Only by embedding care obligations upstream, for example, in design briefs, venture-capital term sheets, and university curricula, can we ensure that shared autonomy serves human flourishing rather than hollowing it out.

Future research should extend this framework to domains beyond mobility and social robotics, including AI-driven hiring platforms that mediate access to livelihoods, algorithmic tutors that reshape childhood learning, and large-language-model assistants that stand between patients and physicians. Each raises its own pattern of dependency and vulnerability, but the diagnostic questions remain the same: who is impacted in what ways, and who remains answerable when things go wrong? A care-centred ethics will not offer a single algorithmic rule; it will, however, keep moral attention fixed where it belongs, on the fragile, interdependent lives that technology should support rather than supplant.

Bibliography

Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I., *The Moral Machine Experiment*, "Nature" 2018, Vol. 563, pp. 59–64.

Bahrami N., *AIgemony: Power Dynamics, Dominant Narratives, and Colonisation*, "AI and Ethics" 2025, Vol. 5, pp. 5081–5103, https://doi.org/10.1007/s43681-025-00734-4.

Beardsley M., Martínez Moreno J., Vujovic M., Santos P., Hernández-Leo D., *Enhancing Consent Forms to Support Participant Decision Making in Multimodal Learning Data Research*, "British Journal of Educational Technology" 2020, Vol. 51, No. 5, pp. 1631–1652, https://doi.org/10.1111/bjet.12983.

Bonnefon J.-F., Shariff A., Rahwan I., *The Trolley, the Bull Bar, and Why Engineers Might Fear Ghosts: An Empirical Study of Morally Loaded Technical Decisions*, "Proceedings of the IEEE" 2019, Vol. 107, No. 3, pp. 502–504, https://doi.org/10.1109/JPROC.2019.2897447.

Bradwell H.L., Winnington R., Thill S., Jones R.B, *Longitudinal Diary Data: Six-Months Real-World Implementation of Affordable Companion Robots for Older People in Supported Living*, in: *Companion Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction*, ACM, New York 2020, pp. 218–220, https://doi.org/10.1145/3371382.3378256.

Clark A., Chalmers D.J., *The Extended Mind*, "Analysis" 1998, Vol. 58, No. 1, pp. 7–19, https://doi.org/10.1093/analys/58.1.7.

European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council*, OJ L 119, URL: https://gdpr-info.eu/.

European Union, *Regulation (EU) 2017/745 of the European Parliament and of the Council*, URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng.

European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, OJ L 1689, 12.07.2024, URL: https://artificialintelligenceact.eu/.

Foot P., *The Problem of Abortion and the Doctrine of Double Effect*, "Oxford Review" 1967, Vol. 5, pp. 5–15.

Geisslinger M., Poszler F., Betz J., Lütge C., Lienkamp M., *Autonomous Driving Ethics: From Trolley Problem to Ethics of Risk*, "Philosophy & Technology" 2021, Vol. 34, No. 4, pp. 1033–1055.

Held V., *The Ethics of Care: Personal, Political, and Global*, Oxford University Press, Oxford 2006.

Jecker N.S., *Nothing to Be Ashamed Of: Sex Robots for Older Adults with Disabilities*, "Journal of Medical Ethics" 2021, Vol. 47, No. 1, pp. 26–32, https://doi.org/10.1136/medethics-2020-106645.

Kittay E.F., *Love's Labor: Essays on Women, Equality, and Dependency*, Routledge, New York 1999.

Lucifora C., Grasso G.M., Perconti P., Plebe A., *Moral Dilemmas in Self-Driving Cars*, "Rivista Internazionale di Filosofia e Psicologia" 2020, Vol. 11, No. 2, pp. 238–250, https://doi.org/10.4453/rifp.2020.0015.

Mackenzie C., Stoljar N., eds., *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, Oxford University Press, New York 2000.

Mele A.R., *Motivation and Agency*, Oxford University Press, Oxford 2003.

National Transportation Safety Board, *Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian*, URL: https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx.

Noddings N., *Caring: A Relational Approach to Ethics and Moral Education*, 2nd ed., University of California Press, Berkeley 2013.

Nyholm S., Smids J., *The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?*, “Ethical Theory and Moral Practice” 2016, Vol. 19, pp. 1275–1289, https://doi.org/10.1007/s10677-016-9745-2.

Oshana M., *Personal Autonomy in Society*, “Journal of Social Philosophy” 1998, Vol. 29, No. 1, pp. 81–102, URL: https://www.academia.edu/download/32898544/Autonomy_and_Society_Journal_of_Social_Philosophy_offprint.pdf.

Pu L., Moyle W., Jones C., Todorovic M., *The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomised Controlled Studies*, “The Gerontologist” 2019, Vol. 59, No. 1, e37–e51, https://doi.org/10.1093/geront/gny046.

Puglisi A., et al., *Social Humanoid Robots for Children with Autism Spectrum Disorder: A Review of Modalities, Indications, and Pitfalls*, “Children” 2022, Vol. 9 , No. 7, 953, https://doi.org/10.3390/children9070953.

Realbotix, URL: https://www.realbotix.com/.

Singleton T., Gerken T., McMahon L., *How a Chatbot Encouraged a Man Who Wanted to Kill the Queen*, BBC News, 6.10.2023, URL: https://www.bbc.com/news/technology-67012224.

Thomson J.J., *Killing, Letting Die, and the Trolley Problem*, “The Monist” 1976, pp. 204–217.

Tronto J., *Caring Democracy: Markets, Equality, and Justice*, New York University Press, New York 2013.

Ullman E., *Programming the Post-Human: Computer Science Redefines “Life”*, “Harper’s Magazine” 2002, Vol. 305(1829), pp. 60–70.

Van de Poel I., *The Problem of Many Hands*, in: I. van de Poel, L. Royakkers, S.D. Zwart, *Moral Responsibility and the Problem of Many Hands*, Routledge, New York 2015, pp. 50–92.

Yamazaki R., Nishio S., Nagata Y., Satake Y., Suzuki M., Kanemoto H., Yamakawa M., Figueroa D., Ishiguro H., Ikeda M., *Long-Term Effect of the Absence of a Companion Robot on Older Adults: A Preliminary Pilot Study*, “Frontiers in Computer Science” 2023, Vol. 5, 1129506, https://doi.org/10.3389/fcomp.2023.1129506.