encyclopedia

MDPI

Entry

Navigating the Ethics of Artificial Intelligence

Jack Harris and Veljko Dubljević *

Science, Technology, and Society (STS) Program, Department of Philosophy and Religious Studies, College of Humanities and Social Sciences, North Carolina State University, Raleigh, NC 27695, USA; jhharri4@ncsu.edu

* Correspondence: veljko_dubljevic@ncsu.edu

Definition

This entry delineates artificial intelligence (AI) ethics and the field's core ethical challenges, surveys the principal normative frameworks in the literature, and offers a historical analysis that traces and explains the shift from ethical monism to ethical pluralism. In particular, it (i) situates the field within the trajectory of AI's technical development, (ii) organizes the field's rationale around challenges regarding alignment, opacity, human oversight, bias and noise, accountability, and questions of agency and patiency, and (iii) compares leading theoretical approaches to address these challenges. We show that AI's development has brought escalating ethical challenges along with a maturation of frameworks proposed to address them. We map an arc from early monisms (e.g., deontology, consequentialism) to a variety of pluralist ethical frameworks (e.g., pluralistic deontology, augmented utilitarianism, moral foundation theory, and the agent-deed-consequence model) alongside pluralist governance regimes (e.g., principles from the Institute of Electrical and Electronics Engineers (IEEE), the United Nations Educational, Scientific and Cultural Organization (UNESCO), and the Asilomar AI principles). We find that pluralism is both normatively and operationally compelling: it mirrors the multidimensional problem space of AI ethics, guards against failures (e.g., reward hacking, emergency exceptions), supports legitimacy across diverse sociotechnical contexts, and coheres with extant principles of AI engineering and governance. Although pluralist models vary in structure and exhibit distinct limitations, when applied with due methodological care, each can furnish a valuable foundation for AI ethics.

Keywords: artificial intelligence; AI ethics; value alignment; black-box problem; human-in-the-loop; ethical pluralism; policy guidelines

Academic Editor: Raffaele Barretta

Received: 17 October 2025
Revised: 10 November 2025
Accepted: 18 November 2025
Published: 26 November 2025

Citation: Harris, J.; Dubljević, V. Navigating the Ethics of Artificial Intelligence. *Encyclopedia* **2025**, *5*, 201. https://doi.org/10.3390/encyclopedia5040201

1. Genesis and History

1.1. On Artificial Intelligence

Artificial Intelligence (AI) refers to computational systems capable of behaviors that humans consider "intelligent," such as learning, reasoning, perception, and problem solving [1]. The field emerged in the mid-20th century, grounded in the aspiration to develop machines capable of emulating human cognitive functions. Early AI scholarship included Alan Turing's 'test' to distinguish between AI and human natural language responses, and a 1956 Dartmouth College workshop that coined the term "AI" [2,3]. John McCarthy, who led the Dartmouth workshop, proceeded "on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [4,5]. With McCarthy's conjecture—now

a goal—in mind, AI research grew markedly over the next seventy years, traversing multiple paradigms.

In the 1950s–1970s, researchers focused on symbolic AI using explicit, rule-based programs. This era emphasized knowledge representation and formal logic (e.g., production rules, theorem provers), such as "expert systems" for diagnostics and decision support [6]. The 1980s–2000s saw the rise of statistical learning methods, including AI utilizing decision trees, support vector machines, and the first multi-layer neural networks. Probabilistic models (e.g., Bayesian networks) and ensemble methods broadened the field, enabling learning from uncertainty and improving generalization [7].

From the 2010s onward, deep learning with neural networks enabled breakthroughs in AI perception and natural language processing [8,9]. These gains were driven by algorithmic advances (e.g., recurrent neural nets, parallel processing), new vast datasets, and accelerated computing capabilities. Most recently, increasingly sophisticated neural architectures (from reinforcement learning systems to large language models, generative AI, and agentic AI) have begun to rival human performance in healthcare, finance, law, and academia, albeit for narrowly defined tasks [10–14].

1.2. Terminological Housekeeping

AI systems have evolved substantially. The earliest symbolic AI encoded knowledge as explicit statements built from logic gates (e.g., "if-then", "and", "or"), and was engineered to apply logical inferences to derive reliable conclusions [9]. AI using decision trees modeled choices not as logic gates, but as hierarchical splits mapping features to predicted outcomes [8,9]. Related AI utilizing support vector machines identified boundaries to split data into groups, and classified data by margin width in a defined feature space [8,9].

Earlier AI systems employing Bayesian probabilistic models were able to represent uncertainty with explicit probabilities. These models updated credences from input data, thereby engaging in a rudimentary form of machine learning [8,9]. Machine learning (ML) refers to systems that are able to learn and fine-tune without being explicitly programmed to do so. ML has become increasingly sophisticated and widespread. AI ensemble systems group and combine ML AI systems that 'boost' or 'stack' each other to improve ML robustness, capacity, and accuracy [8,9].

AI systems utilizing neural networks are comprised of layered function approximators constructed by nodes in an interconnected network. At large scale with many net layers, large datasets, and parallel processing, this is referred to as deep learning [8,9,15]. AI systems using natural language processing (NLP) use algorithms to parse, interpret, and generate human language. Large language models (LLMs) are a class of NLP models that use deep neural networks to execute diverse language functions (such as ChatGPT, Claude, and Gemini).

Generative AI is a broad term for all AI models that synthesize, create, or generate new content such as text, images, audio, or code. Agentic AI denotes ecosystems of AI systems that collectively plan, utilize tools, execute functions, and share memory, which are characteristically able to function without human prompting or oversight [16,17]. They are understood as autonomous or semi-autonomous AI ensembles [17].

AI capacity, architecture, and conceptualization have advanced across symbolic, statistical, and neural paradigms. While the term covers many systems and capabilities, in this entry, AI refers to a broad class of contemporary neural network-based systems, centered on LLMs, generative AI, and agentic AI. While different use cases carry different challenges, this entry consolidates the field's central issues and catalogs ethical frameworks marshaled to address them.

1.3. Structure of the Entry

With the terminological housekeeping now in hand, we start by articulating the rationale for AI ethics. Second, we trace the historical development of leading theories of AI ethics. Third, we compare these theories in the context of contemporary AI systems, highlighting the strengths of pluralist frameworks.

We begin by briefly sketching the novel challenges. Advances in AI have introduced new ethical challenges and magnified longstanding ones. These challenges, among the others outlined in Section 2, motivate the field. First, opacity: the rule-based symbolic systems of half a century ago exposed their reasoning, whereas contemporary generative and agentic models operate as black boxes, insofar as they rely on neural networks, complicating transparency and explainability [18]. Second, data provenance and bias: unlike earlier systems, modern AI models learn from vast, weakly governed datasets that are inscrutable and are prone to encoding biases, heightening auditability and fairness risks. By contrast, AI systems of the 1970s–1980s typically relied on highly curated knowledge bases [9]. Third, autonomy and control: agentic systems can pursue multi-step goals and can independently use external tools, yielding behaviors not explicitly programmed and increasing risks of unpredictability, misuse, and misalignment, concerns far less pronounced in early AI.

The risks are new, and the stakes are significant. AI systems are ubiquitous and are now embedded in many aspects of everyday life. AI systems sort and recommend content on social media, optimize logistics and supply chains, and power virtual assistants on our phones and computers [19]. These technologies influence outcomes of varying magnitude, such as who gets a loan or a job interview, the manner in which resources are allocated, and the way in which people access information [19].

These concerns are not speculative. AI systems have already been found to discriminate in hiring and lending [20], autonomous vehicles have been involved in fatal accidents [21], and chatbots trained on poor data and lacking ethical guardrails have produced misrepresentational outputs [22,23], and even contributed to suicides [24]. AI technologies are being increasingly employed in policing, medicine, law, academia, and warfare. Given AI's rapid development, its permeation across the human experience, and the potentially high stakes of failure, a robust framework for practicing AI ethics is needed.

To guide the responsible design, training, and deployment of AI systems, we need a full-bodied, operationalizable, and explanatorily powerful AI ethics framework. Amid rapid and disorienting change, we need guidance. The remainder of this entry further motivates the rationale for AI ethics, traces the historical emergence of the field, explores major ethical frameworks, illustrates a turn from ethical monism to ethical pluralism, and endorses that progression.

2. The Rationale for AI Ethics

The rationale for AI Ethics can be drawn out through ethical challenges. The most pressing challenges include the alignment problem, black box problem, human-in-the-loop problem, bias and noise problems, issues of accountability, and questions regarding AI agency or patiency. AI Ethics frameworks are best brought out by their reactions to these problems. Each will be glossed in turn.

2.1. The Alignment Problem

One of the most salient ethical challenges in AI is the alignment problem: ensuring that AI systems pursue goals that are consistent with human values and intentions. Following Nick Bostrom's influential formulation of the alignment problem in *Superintelligence: Paths, Dangers, Strategies*, particularly his analysis of the challenges of control and the specification

of human values, alignment has become a central organizing concept in AI ethics [25]. The term was further popularized by theorists such as Brian Christian, Leonard Dung, and Stuart Russel [22,23,26]. As Leonard Dung asks, "How can we build AI systems such that they try to do what we want them to do? This, in a nutshell, is the alignment problem" [23]. Alignment is difficult both in terms of value setting (deciding what values or objectives AI should have) and value implementation (successfully embedding those values into machines) [27]. Different ethical theories take different positions on both elements of alignment.

The first prong of the problem probes: which ethical theory should we use to align AI? There are many contenders (Table 1), but each faces challenges. The second prong of the problem asks: how should we represent and operationalize ethical principles in a machine-readable format? How can we prevent phenomena like alignment drift (an AI's goals shifting over time) or alignment faking (an AI appearing aligned while concealing its true objectives)? Many theorists warn that misaligned AI could lead to catastrophic outcomes [28–31]. Some theorists distinguish between the broader goal of beneficial AI (well-aligned value sets) versus technical alignment (effectively implemented values), with many prioritizing the former [32,33]. Even if an AI faithfully optimizes for a given value, that value itself might be poorly chosen. Setting values necessitates an ethical framework. Specifying how AI design should internalize and reflect values has driven the application of a variety of ethical frameworks, each a candidate for alignment.

Table 1. Ethical Theories and AI Alignment.

	Deontological Monism	**Deontological Pluralism**	**Consequentialism**	**Virtue Ethics**	**Contractualism**	**Agent-Deed-Consequence**
Value Setting	Simple	Simple/ Complex	Simple	Complex	Complex	Complex
Value Implementation	Simple	Complex	Simple	Complex	Simple	Simple/Complex
Breadth	Narrow	Wide	Narrow	Narrow/Wide	Narrow/Wide	Very Wide
Rigidity	High	Low	High	Low	Low/High	Low

The alignment problem matters to ethicists and engineers alike. AI engineers must recognize that they are not creating systems from a neutral "view from nowhere" [34]—their design choices implicitly embed values and priorities. If we follow Cathy O'Neil's reasoning that "algorithms are opinions embedded in mathematics", then AI is inherently value-laden [35]. If this is the case, stakeholders must consider which alignment strategy is best, along with how best to implement it. Alignment might involve hard-coding moral principles (a deontological approach), training AI with reward functions reflecting ethical utilities (a consequentialist approach), or a variety of hybrid views. Another controversy concerns the standpoint relative to which AI systems ought to be aligned: while most accounts presuppose an anthropocentric orientation, alternative approaches have emerged, such as Banerjee's "cosmicist" view, which suggests alignment criteria that decenter human interests in favor of a stance of cosmic humility [36]. Each approach towards alignment has limitations of breadth and rigidity.

2.2. The Black Box Problem

Another ethical challenge for AI is the so-called "black box" problem of opacity [37]. The black box problem applies to neural networks, "sophisticated self-learning forms that continuously test and adapt their own analysis procedures" [38]. Neural networks are parameterized function approximators composed of layers of interconnected nodes. Interior nodes of neural networks transform input data through weighted linear combinations and nonlinear activation functions [39]. These interior nodes are capable of revising their semantic content and their relations to other nodes, and do so iteratively until they

produce an output. Neural networks are vital to many AI applications, including AI vision, generative AI, and predictive analytics [40]. Neural networks may model dimensionally rich patterns analogously to human synaptic transmission.

That power, however, entails costs. While input and output nodes are observable, scrutable, and clear, interior nodes are not. Rather, they are said to be 'black boxed' [41–43]—impenetrable and unobservable. Since neural networks can revise the value of each node, the semantic content of each node, and the arrows of influence between each node (bidirectionally, iteratively, and with massive speed and scale), it is effectively impossible for humans to trace how inputs are transformed into outputs [43].

Assessments differ on the severity of the black-box problem. Verdicts vary on whether neural networks are intractably opaque and the extent to which opacity is problematic. Some theorists think that we can gain insight into the interior node content and parameterization of a neural net, for instance, by attribution mapping, node or neuron whitening/white boxing, or counterfactual tests [41–43]. Other theorists doubt this [43]. Alternatively, some downplay the importance of opacity, arguing that strong model assurance, or "computational reliabilism," makes black box worries less problematic [37], but most regard opacity as a serious issue.

Epistemic opacity was a non-issue for early symbolic or "good old-fashioned" AI (GOFAI), whose rule-based representations and inference chains were explicit and auditable [43]. Earlier AI architectures implemented hand-specified rules upon discrete symbolic representations, making their internal operations and derivation steps relatively transparent. Developed within the "physical symbol system" paradigm of Newell and Simon [44], such systems were therefore far less susceptible to the forms of epistemic opacity that now characterize black-box models. By contrast, the black-box problem is acute for contemporary systems. This matters not only because it frustrates scientific understanding, but because many governance aims (explaining adverse outcomes, offering user recourse, detecting bias, and allocating responsibility) presuppose explanatory access to the causal basis of a system's behavior.

The ethical importance of explainability is clear, as seen through the theories canvassed in Section 3. From a Kantian deontological perspective, opaque AI systems threaten autonomy because users and human inputs cannot consent to or understand decisions affecting them. From a utilitarian consequentialist perspective, opacity makes it harder to predict consequences and correct errors, potentially reducing overall welfare, and challenging the architecture of 'rule utilitarian' policies. Contractualist approaches demand that decision processes be justifiable to those whom they affect. Pluralistic frameworks may object that without explainability, we cannot assess considerations of virtue, consent, or autonomy. Opacity also complicates accountability and governance: we may not be equipped to govern or regulate what we do not understand. Ethical frameworks differ in how they problematize and assess the black-box problem. Its growing threat underscores the need for further conceptual convergence in AI ethics.

2.3. The Human in the Loop Problem

The "human in the loop" problem mandates that decisions supported or executed by AI must be overseen by humans with genuine intervention power. Framed as a response to AI's limits (including opacity), the problem demands meaningful human oversight and control, either as a cooperative team member, a safety backstop, or an adversarial challenger. Whether we can (or should) integrate meaningful human oversight and control over an AI is a matter of substantial debate [45]. If an AI's internal logic is a black box and its objectives may misalign, having human judgment as a check can prevent or catch

dangerous errors, but human intervention can also introduce bias, inconsistency, and new forms of moral and legal risk [46].

In practice, humans can be kept "in the loop" at various stages. Upstream, humans can curate training data, design algorithms, and set goals before an AI system is deployed [45,46]. Downstream, humans can review or override AI outputs during operation; for instance, a physician may confirm or veto an AI-generated diagnosis prior to issuing it, or a driver may take over from an autonomous vehicle at any time [47]. A life-cycle approach integrates human oversight throughout the AI system's design, deployment, and maintenance.

The life-cycle model may be ideal in theory, but it comes with costs. It can be difficult in practice due to cost and complexity, and it may be impossible with neural network algorithms (as humans cannot intervene upon the impenetrable layers of the network), or frontier developments like agentic AI (as humans cannot effectively intervene upon "multi-agentic ecosystems") [17,47]. There are also concerns that excessive human intervention could reintroduce biases or inefficiencies that AI was meant to reduce, or could sacrifice the scalability of AI systems, which are a substantial boon.

Determining whether and where human oversight is indispensable is an ongoing challenge. In domains like autonomous weapons or self-driving cars, there are strong reasons to insist on human intervention before life-and-death decisions are made [48]. In other areas, we may decide that properly validated AI can be trusted to operate autonomously within set bounds. Wherever it lies, the key is to strike an appropriate balance: leveraging AI's efficiency, scalability, and consistency while preserving human values and common-sense judgment to handle the cases or conflicts that AI cannot, and to intervene when AI has gone awry. Different ethical theories reach different conclusions about the need for human-in-the-loop governance.

2.4. Fairness, Bias, and Noise

Although AI systems are often marketed as objective decision-makers [34], they can introduce, reproduce, or amplify bias and noise. Output quality is bounded by input quality: pretraining data that is incomplete, low-resolution, noisy, or unrepresentative leads to biased or unreliable models. Noise (stochastic variability) and bias (patterned variability) can enter at multiple points in the process of developing an AI system: in data framing, data collection, measurement and labeling (e.g., annotator bias), modeling choices, and deployment (e.g., feedback loops and distributional shift) [49].

Theorists have long emphasized that no algorithm is truly "neutral". High-profile incidents illustrate the point: in 2018, Amazon found that an experimental AI hiring tool was implicitly discriminating against women because it had learned from historical hiring data dominated by male candidates [21]. Facial recognition systems have likewise exhibited markedly higher error rates for people with darker skin, traceable to underrepresentation and labeling biases in training data [50]. Divergence is not only the by-product of randomness or error; it is often structural, arising from designs that target specific objectives, organizational aims, or value assumptions.

Though structural AI biases exist, many biases are unintended and undesired. A core challenge is how to detect, quantify, and mitigate bias across the AI model's lifecycle [51]. Potentially effective measures include rigorous data governance and curation, bias or drift auditing, and the adoption of FAIR (findable, accessible, interoperable, reusable) data principles [52].

Aside from bias, AI systems can also suffer from noise, random variability in judgment, outcome, or output caused by externalities that ought not to influence a decision [53]. Human decision-makers are notoriously noisy; significant documented inter-judge or inter-clinician variance evidence this [53]. Replacing human decision makers with AI can

reduce noisy idiosyncratic variance. Indeed, Cass Sunstein has argued that governing by algorithm produces "no noise and (potentially) less bias" [49,54]. However, other theorists have cautioned that AI introduces its own noise sources: small input perturbations or preprocessing differences, potentially stochastic inference procedures (e.g., sampling, dropout), and distributional drift can yield different outcomes for near-identical cases [55]. Other theorists have stressed that we ought to develop "noise-aware" artificial intelligence systems; rather than taking AI to be invulnerable to noise, perhaps AI had better flag it [56]. Unchecked, AI noise can be more pernicious than AI bias because it is harder to detect, target, audit, and contest.

2.5. Accountability and Responsibility

When AI systems cause harm, make mistakes, or produce desirable outcomes, it is often unclear who or what should be held accountable. Traditional notions of responsibility become muddled: engineers who built the AI, the company that deployed it, the end-users, the data providers, and even the AI itself (though an AI is not a legal person) could all be seen as partly responsible. Because AI decision-making is distributed across many actors and components, failures tend to result from a chain of events rather than a single identifiable choice [57].

This diffusion of agency makes it challenging to assign blame, liability, or praise. Agency is not only distributed between agents (engineers, companies, and users), but it is also distributed within an AI system. The latter diffusion has given rise to what is known as the "credit assignment problem": the difficulty of figuring out which element (particular data sources, model components, design choices, nodes, or node connections) deserves blame (targeting) or praise (reinforcement) for a given outcome [58,59]. The credit assignment problem predates contemporary AI systems and was problematic for earlier symbolic AI of the 1980s [59]. The problem, however, has gotten much harder to solve with the advent of neural networks and generative AI, because outputs arise from distributed components and temporally extended learning, causally linking an outcome back to the responsible state, action, or actor is exceedingly difficult.

So-called algorithmic accountability is challenging from all angles. Consider an accident involving a self-driving car [60]. Who is responsible for the crash? The possibilities include the car's owner, the vehicle manufacturer, the software developers, a third-party data provider, the autonomous system itself, or specific fine-grained elements of the autonomous system (sensors, decision procedures, motor devices). This is problematized further with agentic AI systems, which operate as autonomous ecosystems of cooperative agents, capable of acting autonomously as a collective [61].

Currently, legal systems do not recognize AI as bearing responsibility, so fault must be traced back to human or corporate agents. This has led to debates and proposals around updating laws and norms pertaining to AI. Some suggest that every AI system should have an identified human supervisor or operator who is accountable for its actions [62,63]. Others have argued for expanding product liability to software, treating an AI error like a defective product for which the manufacturer is liable [64]. More radically, some scholars have proposed a form of legal personhood for advanced AI agents, allowing them to be, for instance, sued or insured in their own right [64]. Although consensus is still emerging, a widely shared premise is that AI systems require determinate accountability; without it, we risk a moral and legal vacuum. Clarifying the scope of human responsibility is therefore essential to avoid responsibility gaps. Which agents (or artifacts) are answerable, and on what grounds, depends on the ethical framework adopted, strengthening the case for robust, unified, and better-theorized AI ethics.

2.6. AI Agents, Patients, and Personhood

AI ethics also ventures into questions of metaphysics, asking what sort of entities AI systems are (or could become) in moral terms. Fundamental questions here include: can an AI be a moral agent (capable of making ethical decisions and being held responsible for them)? Could an AI be a moral patient (deserving of moral consideration and possibly rights, if it has experiences or consciousness)? These questions, once purely speculative, are starting to be taken seriously. Some AI developers, such as those at the Anthropic corporation, have even begun exploring and operationalizing AI welfare [65,66].

Early insights regarding AI agency came from Luciano Floridi and J.W. Sanders' 2004 work on the moral status of artificial agents [67]. They argue that AI systems extend the class of entities involved in ethical situations because they can function both as moral patients (entities that are acted upon for good or ill) and as moral agents (entities that perform actions with broadly moral impacts) [67]. Crucially, they draw a distinction between an agent causing an outcome (accountability) and an agent being blameworthy for that outcome (responsibility); the two are not coextensive. Using this distinction, they explain how an artificial agent can be a moral agent capable of producing harm or benefit without being a morally responsible agent in the human sense. In other words, an AI might be the immediate cause of an accident or decision (making it accountable in a causal sense), but we do not hold the AI itself culpable or blame the algorithm for the outcome; this is what they refer to as "mind-less morality," meaning moral agency does not require consciousness or free will at a given level of abstraction.

On Floridi and Sanders' account, the moral and legal responsibility for an AI's actions must still be traced back to human designers, operators, or institutions, since current AIs cannot understand or answer to moral blame, nor can they be held liable in a court of law [67]. Similar positions have been echoed by theorists, including Caplan and colleagues [68]. To prevent ethical and legal gaps in complex scenarios where agency is distributed across many human and AI actors, Floridi's later work on "distributed morality" proposes assigning "faultless responsibility" to all causally relevant nodes in an AI-driven network by default, essentially a strict-liability-inspired approach so that accountability is not lost amid distributed decision-making and challenges of credit assignment [69].

David Chalmers argues that we should prepare for the possibility that AI systems may exemplify significant elements of personhood and consciousness in the near future, positing a "substrate-independent" view of these features [70]. He suggests that institutions should establish procedures for treating AI systems with appropriate moral concern. In a similar vein, Robert Long and colleagues (Chalmers among them) identify three strategies that institutions can implement to prepare for the potential emergence of AI consciousness: "They can (1) acknowledge that AI welfare is an important and difficult issue (and ensure that language model outputs do the same), (2) start assessing AI systems for evidence of consciousness and robust agency, and (3) prepare policies and procedures for treating AI systems with an appropriate level of moral concern" [71]. Of course, questions remain to be filled in. What would "an appropriate level of moral concern" actually look like [72,73]? According to Kyle Fish, Anthropic's AI welfare researcher, it could take the form of allowing an AI model to stop a conversation with a human if the conversation turned abusive: "If a user is persistently requesting harmful content despite the model's refusals and attempts at redirection, could we allow the model simply to end that interaction?" [66]. As of August 2025, Anthropic's Claude chatbot can now unilaterally end conversations with users and can object to perceived threats [74].

Even short of achieving personhood and entitlement to patiency responses, AI systems are increasingly functioning in roles that invite human empathy or trust (such as caregiver robots or virtual assistants with personalities). Evidence indicates that some individuals

form relational ties with their AI systems. A debate has arisen over whether these systems should be granted any kind of moral or legal status. A small number of theorists argue that as AI systems integrate into social roles, we might consider granting them certain protections or status, such as rights [70,71]. On the other hand, others caution against anthropomorphizing AI. Joanna Bryson provocatively stated that "robots should be slaves;" indeed, the etymological roots of the term robot are literally 'slave', first introduced in Karel Čapek's 1920 play "Rossum's Universal Robots" (robot is derived from the Czech word *robota* [forced work]) [75,76]. Bryson, among other theorists, maintains that granting rights to current AIs is misguided and dangerous: it could dilute human rights, disempower human rightsholders, and provide a convenient excuse for human decision-makers to avoid responsibility [75].

Some theorists see AI as a potential agent and rightsholder; others take it to be a tool. Current debates lack an adequate taxonomy for entities that straddle the agent/tool divide, such as displaying ward-like, aspirant, or apprentice forms of agency that act with constrained autonomy and require structured oversight. Perhaps AI systems ought to be teamed with, or treated as a journeyman, as has been recently suggested [77]. Finally, emerging literature proposes modeling agentic AI as group agents, which raises unsettled issues about decision procedures, cooperation, and group agency [17]. These agent and patient-based gray zones are undertheorized and merit ethical attention.

The rationale for AI ethics is well-defined. Only armed with an ethical theory can one tackle problems of AI alignment, opacity, limits on human oversight, bias and noise, diffuse accountability, and potential agency. Addressing these challenges requires defensible value choices, coherent normative reasons, and criticizable decision procedures. Ethical frameworks offer precisely these resources.

3. An Arc Towards Pluralism in AI Ethics

This section maps the historical evolution of AI ethics (Table 2). Categorizing AI ethics frameworks is not a merely historical exercise, but a way of clarifying the normative assumptions that underwrite contemporary principles, guidelines, and governance tools. Distinguishing between monistic and pluralistic, deontological and consequentialist, and hybrid approaches makes visible how different theories identify, prioritize, and trade off values such as welfare, autonomy, justice, and accountability, and reveals how they generate distinct design and regulatory prescriptions.

Table 2. A Historical Survey of Theories of AI Ethics.

Date—Theorist	Theory Type	Key Theory Elements
1950—Asimov	Deontological Monism	Rule-based AI ethics [78].
1955—Wiener	Consequentialist Monism	AI for social good, broadly utilitarian [79,80].
1976—Weizenbaum	Deontological Monism	Duties predicated on autonomy and justice pertaining to AI use [81,82].
1976—Maner	Consequentialist Monism	Classical utilitarianism [83].
1979—Beauchamp & Childress	Deontological Pluralism	Four non-absolute principles [84,85].
1979—Moor	Hybrid	"Just consequentialism", consequentialism after justice constraints [86–88].
1990—Clouser & Gert	Anti-Pluralism	Critiquing pluralism in applied ethics, implications for AI ethics [89].
2005—Anderson & Anderson	Deontological Pluralism	Rossian AI ethics duties, operationalized in a machine-readable format [90–92].
2008—Wallach & Allen	Hybrid Pluralism	Implementable pluralism for moral machines using abductive reasoning [93].
2014—Dubljević & Racine	Hybrid Pluralism	ADC meta-pluralism about normative sources [94].
2022—Telkamp and Anderson	Hybrid Pluralism	Moral foundation theory, six irreducible descriptive normative sources [95].
2024—Gros, Kester, Martens, and Werkhoven	Consequentialist Pluralism	Augmented utilitarian principles [96].

Tracing the field's movement from early monistic models toward increasingly pluralistic and hybrid frameworks, therefore, serves two purposes: it reveals an arc of conceptual maturation in response to the multidimensional challenges of AI, and it provides a taxonomy for evaluating which families of theories are best positioned to structure robust and operationalizable solutions for AI ethics.

3.1. A Historical Survey of Theories of AI Ethics

3.1.1. Asimov—Early Deontology

Frameworks in AI ethics, which structure responses to the problems just outlined, have evolved significantly from the mid-20th century. In the early decades of AI (1950s–1970s), explicit discussion of "AI ethics" was sparse and mostly appeared in speculative contexts like science fiction. A famous example is Isaac Asimov's Three Laws of Robotics, introduced in a 1942 short story and later popularized in his book *I, Robot* (1950). These fictional laws, foremost that "a robot may not injure a human being," were an early attempt to imagine built-in ethical safeguards for machines with artificial intelligence [78]. This is an early example of monistic deontology: a rule-based framework for AI ethics, with a single overarching rule or decision procedure.

3.1.2. Wiener and Maner—Early Consequentialism

Other early approaches included utilitarianism, a form of monistic consequentialism. In 1950, Norbert Wiener framed automation and "ultra-rapid computing" in terms of impacts on human welfare [79]. Wiener is credited with having established one of the first comprehensive "information ethics" frameworks [80]. Wiener's framework is fundamentally consequence-sensitive, foregrounding risks of unemployment, exploitation, deleterious feedback effects of automation, and long-run social instability. However, he explicitly rejects any monistic maximizing metric (such as efficiency, profit, or aggregate utility), instead advancing a nuanced form of consequentialist reasoning aimed at broad welfare.

Following Wiener, Walter Maner's "computer ethics" framework adopted a similar approach. In the mid-1970s, Maner's theory, which explicitly recommended applying classical utilitarian analysis to "computer ethics" (a term which he coined), gained considerable traction [81,83]. Maner treated the field of computer ethics as warranted by the distinctive negative consequences (harms and risks) produced or transformed by computing technologies. Deontology and consequentialism were the first moral frameworks on offer, owing to their ease, simplicity, and formulaic nature.

3.1.3. Weizenbaum—Accommodating Autonomy and Justice

Concerns about justice, autonomy, and transparency, poorly handled by earlier monistic theories, soon accompanied advances in AI. In 1976, Joseph Weizenbaum, creator of the pioneering *ELIZA* chatbot (considered to be the world's first), argued in *Computer Power and Human Reason* [82] that reliance on machines must be bounded: responsibilities that demand empathy, compassion, practical judgment, or the exercise of one's own autonomy should not be delegated to AI, even when deontological rule-following may be plausible, or desirable outcomes might be produced. Neither deontological nor utilitarian theory adequately addressed these concerns. Weizenbaum called for clear, principle-based limits grounded in autonomy, justice, and care. This marked an early turn toward pluralism.

3.1.4. Beauchamp and Childress—Deontological Pluralism

Building on this pluralistic trajectory, and the 1978 Belmont Report [84], Tom Beauchamp (a principal drafter of Belmont) and James Childress developed a pluralistic framework for biomedical ethics that has greatly influenced AI ethics. In *Principles of*

Biomedical Ethics [85], they advanced four general moral principles: respect for autonomy, beneficence, non-maleficence, and justice. These principles are grounded in the deontological pluralism of W. D. Ross [90] as a common framework for ethical decision-making. Crucially, these principles are *prima facie* duties: each is morally binding unless it conflicts with another, in which case specification, balancing, and judgment are required. This approach is explicitly pluralistic: rather than a single absolute rule, it offers multiple non-absolute principles to guide action. Beauchamp and Childress deliberately drew on Kantian duty (autonomy), utilitarian considerations (beneficence, non-maleficence), and diverse theories of justice (egalitarian, liberal-egalitarian, libertarian) to avoid commitment to any one dominant theory. Their pluralist approach gained prominence in medicine and has informed principle-based frameworks in AI ethics; for a recent and influential account inspired by Beauchamp and Childress, see Floridi and Cowls, Cortese et al., and Adams [97–99]. Now-familiar claims that ethical AI should avoid harm, promote good, respect autonomy, and attend to justice trace back to this model of balancing duties.

3.1.5. Moor—Just Consequentialism

James Moor advanced an early pluralist framework for AI and computer ethics in 1979, though it was distinct from Beauchamp and Childress's principlism [86–88]. Moor proposed "just consequentialism": under this view, one imposes deontological constraints of justice (largely from rights), then selects, among the permissible options, the one with the best consequences. In other words, justice- and rights-based constraints operate as screening conditions, and only within the remaining set of permissible acts does consequentialist ranking identify the option with the greatest expected value, thus mitigating utilitarian shortcomings while preserving consequence sensitivity. Like previous theorists, Moor held that some decisions must remain human, as they are ruled out by these constraints [88]. Moor's later work warned against ceding excessive power to AI, while reinforcing the need for a pluralistic mode of ethical assessment.

3.1.6. Clouser and Gert—Pluralistic Skepticism

Not everyone embraced pluralism. In 1990, K. D. Clouser and Bernard Gert published "A Critique of Principlism," arguing that the Beauchamp and Childress (or Rossian) approach to ethics was too vague and unsystematic [89]. Listing principles, they argued, provides little guidance for adjudicating conflicts among them. At best, such principles function as a checklist; at worst, they muddle moral reasoning by mixing heterogeneous theories. Although framed within biomedical ethics, their critique applies with equal force to AI ethics, which likewise relies on high-level principle lists (e.g., beneficence, non-maleficence, autonomy, justice, explicability) without a shared weighing procedure. They warned that ethical pluralism needs a unifying method or theory or else it risks becoming an "empty" exercise, based on a messy "heap" of duties [89]. This criticism laid the groundwork for contemporary attempts to make AI ethics (and pluralistic frameworks generally) more cohesive, better unified, and operationalizable.

3.1.7. Anderson and Anderson—Embodied Pluralistic Deontology

By the mid-2000s, attention shifted from merely designing ethical AI to building moral reasoning into systems [92]. Michael and Susan Leigh Anderson advanced this "machine ethics" agenda by using Ross's pluralistic deontology (at this point, already influential in bioethics) as a blueprint for AI. They implemented a prototype medical-ethics advisor that encoded multiple *prima facie* duties and balanced them when in conflict (e.g., truth-telling vs. non-maleficence), rather than applying a single rule [90]. This demonstrated that pluralism can be operationalized in software for context-sensitive decisions. Pluralistic deontology,

adapted from biomedical ethics, marked a turn toward pluralism; yet, as a duty-based approach grounded in antecedently specified *prima facie* obligations, it remained top-down.

3.1.8. Wallach and Allen—Abductive Hybrid Pluralism

The next advance deepened pluralism by coupling it with abductive reasoning rather than deduction or induction. Wendell Wallach and Colin Allen analyzed how to embed and implement moral reasoning capacities in a machine-readable format, following the work of Anderson and Anderson. Wallach and Allen side with pluralism, contrasting top-down pluralistic approaches (with explicit rules or principles) with bottom-up pluralistic approaches (norms that emerge from outputs), and instead argue for an abductive hybrid approach integrating both strategies [93]. Pure rules are brittle and may be disconnected from use-cases; pure results-based learning is difficult to anticipate and misalignment-prone. A hybrid approach equips agents with guiding principles while allowing adaptation from feedback, each mechanism checking and informing the other. Since the mid-2000s, pluralism coupled with abductive methodology has moved to the fore in AI ethics. This new family of views joins rule-based guardrails with outcome-sensitive reasoning [93].

3.1.9. Dubljević and Racine—Agent Deed Consequence

In 2014, Veljko Dubljević and Eric Racine introduced the Agent–Deed–Consequence (ADC) model, a novel framework integrating insights from virtue ethics, deontology, and consequentialism [94,100–105]. This is a form of meta-pluralism, expanding the bounds of moral consideration beyond the deontology of Ross, or Beauchamp and Childress. Rather than limiting moral evaluation to one perspective (e.g., deontological rules or consequences), the ADC model posits that human moral judgments naturally involve three intuitive components: our perception of the agent's character or intent (Agent), the inherent rightness or wrongness of the action itself (Deed), and the outcomes produced (Consequence) [94]. In any moral inquiry, we simultaneously ask (and ought to ask): was the agent well-intentioned; were fitting agential characteristics displayed? Was the act itself permissible qua the rules on the table? And did it lead to good or bad results? For example, if an AI performs an action that breaks a rule (a negative deed), we might forgive it if the AI's goal was compassionate (a praiseworthy agent) and the outcome turned out to be beneficial (a desirable consequence). This hybrid pluralism mirrors how people reconcile conflicting considerations.

The ADC framework has been proposed as a guide for AI ethics because it provides a flexible, human-like approach to moral reasoning that is faithful to moral phenomenology. Recently, theorists have explored how to implement the ADC model in AI decision systems [94,100–105], reflecting a broader turn toward capturing the complexity of moral judgment in AI design, assessment, and governance. ADC is a meta-pluralistic framework because it evaluates three independent yet interacting moral dimensions at once. This structure preserves value-pluralism, surfaces distinct failure modes on each ethical axis, and supports principled trade-offs rather than collapsing everything into a single metric (e.g., duty-weight, as under pluralistic deontology). The ADC model, at present, is at the forefront of the field's progression toward normative pluralism. Challenges include cross-axis weighting, operationalizing virtues, measuring long-term effects under uncertainty, and systematizing the principles; nevertheless, recent scholarship is rendering these problems tractable (for instance, systematizing ADC procedures by considering agential characteristics in the context of AI scope and governance, deed-constraints in the context of AI design controls, and consequences in the evaluation of relevant AI system objectives) [94,100–105].

3.1.10. Telkamp and Anderson—Moral Foundation Theory

In 2022, Jake Telkamp and Marc Anderson argued that moral foundation theory (MFT) from social psychology should guide AI ethics. It holds that the ethical assessment of AI systems will be irreducibly multi-valued because people weigh six distinct moral foundations: care, fairness, loyalty, authority, purity, and liberty [95]. It may be framed as an objective list theory, though Telkamp and Anderson take this to be descriptive pluralism, rather than prescriptive. Under this theory, all conflicts of principles or duties can be mapped to foundation conflicts (e.g., of fairness vs. liberty). Though seemingly akin to pluralistic deontology, MFT commits neither to deontology nor consequentialism [95]. It is characterized as descriptive intuitionism. Using MFT can better bring into focus ethically fraught tradeoffs involving AI. For instance, AI-based surveillance may promote authority at the cost of liberty. All six moral foundations produce morally salient, but often divergent, responses [95].

By way of illustration, in ethically designing an autonomous vehicle crash policy, care would prioritize minimizing expected injuries; loyalty, protecting occupants; authority, obeying traffic laws and signals; fairness, avoiding discriminatory treatment (e.g., no age- or status-based weighting); purity, avoiding clearly reckless maneuvers; and liberty, respecting others' autonomy. For Telkamp and Anderson, disagreements are predictable, not aberrant, and occur across varied domains: the organizational use of an AI system, the data input into an AI system, and the outcomes produced by an AI system. Each dimension is likely to engage different moral foundations and generate different trade-offs [95].

3.1.11. Gros, Kester, Martens, and Werkhoven—Augmented Utilitarianism

In 2024, Chloe Gros, Leon Kester, Marieke Martens, and Peter Werkhoven argued that augmented utilitarianism should underlie AI ethics. Augmented utilitarianism (AU), like ADC and MFT, is pluralistic [96]. The theory was originally posited to counter reward hacking or 'perverse' goal fulfillment. AU proposes a utilitarian foundation (namely of harm minimization, rather than benefit maximization) that builds an ethical goal function to govern AI by integrating insights from consequentialism, deontology, virtue ethics and other theories [96]. These insights allow the utilitarian framework to capture dyadic harm: harm as an emergent property arising from situated relations between an agent, an act, and a victim [96]. AU, though harm-centric, accommodates a sufficiently broad category of harm to count as pluralistic, including breaches of duty, viciousness, unjust infringements of contractualist principles, human rights violations, and other considerations not neatly captured by monistic consequentialism.

For instance, in ethically considering an AI-based hiring platform under AU, parties should aim to design a harm-minimizing goal function that the model must optimize. They would first have to identify the core consequentialist harms, such as wrongful rejections and wrongful acceptances. Then, harm-based constraints from other theories would be accounted for: non-discrimination across protected groups (fairness), respect for applicant privacy/consent (Kantian deontology), and the requirement to provide justificatory reasons on request (transparency). The AI hiring platform's goal function should operate on this basis, and the AI system should be morally assessed along these lines, under the auspices of AU [96].

3.1.12. Pluralism in View

Pluralistic deontology shows that multiple duties can be encoded and balanced in AI systems, and ought to be used to govern them. Hybrid, abductive approaches join top-down principles with bottom-up learning, so rules and results constrain each other. The ADC

model widens evaluation to agential, deed-constraint, and consequentialist considerations, preserving value pluralism and exposing distinct failure modes. MFT supplies a descriptive map of why stakeholders disagree and how to design procedurally legitimate processes. AU offers a harm-minimizing goal function that incorporates constraints and virtues from other theories to avoid perverse optimization. While hybrid abductive pluralism has gained notable momentum over the last decade, no framework is cost-free; Section 3.2 compares the principal benefits and liabilities of each.

3.1.13. Pluralism in Practice

Recent policy developments reinforced the turn toward pluralism. From the mid-2010s, research and policy communities issued high-level guidance consonant with pluralistic ethics. The Beneficial AI conference at Asilomar (2017) produced 23 principles spanning research priorities, professional ethics, and long-term risk [106]; though intentionally broad, these tenets (including safety, transparency, value-alignment, and social benefit) garnered wide endorsement and shaped later initiatives [107]. In parallel, IEEE's Ethically Aligned Design (EAD) effort released an initial framework in 2016 and a comprehensive first edition in 2019 for autonomous and intelligent systems, emphasizing human rights, well-being, responsibility, transparency, and accountability [108,109]. Together, Asilomar and IEEE EAD exemplify hybrid pluralism: multiple salient values aggregated, with none elevated as supreme. Aligning with this, in 2021, UNESCO translated a pluralist consensus into a globally endorsed policy package adopted by 193 member states, coupling human-rights commitments with standards for data governance, AI education, and appraising the environmental impact of AI systems [110]. The breadth of the Asilomar, IEEE, and UNESCO encoded values most closely mirror the ADC, MFT and AU frameworks, which are among widest forms of pluralism.

3.2. Assessing Frameworks

AI ethics frameworks offer distinct features, insights, and trade-offs. A useful high-level distinction is between monist theories, which elevate a single foundational value or rule, and pluralist theories, which recognize several co-equal moral principles; in this domain, utilitarianism (consequentialism) and Kantianism (deontology) exemplify monism, whereas deontological and consequentialist pluralism and hybrid models are pluralist. Having descriptively traced the field's shift toward pluralism, we now offer a prescriptive case for why this shift constitutes progress for AI ethics as a field.

3.2.1. Consequentialism

Consequentialist theories assess actions by their outcomes. Classical utilitarianism, developed by Jeremy Bentham, John Stuart Mill, and Henry Sidgwick, holds that the right action maximizes overall well-being, happiness, or desirable states of consciousness [111–113]. In AI, a utilitarian approach would appraise and design systems based on the extent to which they optimize aggregate results (e.g., saving the most lives, maximizing efficiency, increasing profit, and so on), a posture that aligns naturally with metric-driven engineering and cost–benefit analysis. Historically, early AI applications often embodied straightforward utilitarian aims: mid-20th-century game-playing programs were built simply to win (e.g., Dietrich Prinz's University of Manchester chess projects) [114]; early AI-assisted trading optimized returns under profit-and-loss criteria; AI systems power at minimum 60% of all stock trading volume in the United States [115].

Though fruitful and simple, utilitarianism faces familiar objections. It can license rights violations or unfair treatment if these raise net utility; an outcome-maximizing system may, for example, deceive or manipulate users to improve metrics, or may engage in reward function gaming [116]. Moreover, calculating the "greatest good" is epistemically

fragile in practice, as it requires prediction of complex downstream effects (which may be very difficult to foresee, particularly in AI use cases). Finally, utilitarianism may require contentious interpersonal comparison and aggregation, raising concerns about the separateness of persons [116]. Accordingly, while utilitarian principles can productively guide certain moral calculations, they should be bound by complementary ethical constraints to avoid serious moral failures.

3.2.2. Monistic Deontology

Across the aisle from utilitarianism, deontological ethics focuses on duties or rules rather than consequences, on input rather than output. Monistic deontology posits a single overriding duty. Kantian deontology is particularly well known; such an approach to AI ethics would insist that certain moral rules (such as respecting human autonomy and dignity by treating agents as ends in themselves) must never be violated, no matter the potential benefit [117]. In practice, this might mean building hard constraints into AI systems. For example, under these auspices, an AI should never deceive or coerce a human because that would treat the person as a mere means to an end (robbing them of a chance to exercise their will, and to "contain in themselves the end of an action," violating their autonomy) [117]. The strength of a deontological monist approach is that it provides firm ethical guardrails. It protects fundamental rights and values from being overridden by calculations of utility [118]. Many proposals in AI ethics echo Kantian themes: for example, requirements that systems secure informed consent for certain actions; principles mandating intelligibility and meaningful human control; and prohibitions on deceptive or coercive design [118,119].

The limitation, however, is that strict rules can lead to moral rigidity. Cases often involve conflicting duties, and a hardline rule-based system struggles when two imperatives collide. Suppose that an AI assistant designed via monistic deontology has a hard rule: never lie to the user. What if the user asks a question whose truthful answer is likely to trigger harm? Consider Jonathan Rinderknecht, the suspect accused of starting the deadly Palisades Fire in Los Angeles, who asked ChatGPT, "Are you at fault if a fire is lit because of your cigarettes?" [120]. Should the AI system reveal potential liability gaps that may enable one to start such a fire? Deontology has a hard time admitting to exceptions. Moreover, there is indeterminacy as to which rule should be paramount: not everyone agrees on a single supreme moral principle, which makes monistic approaches highly contentious. Is the top priority to respect autonomy, to ensure equity, to minimize harm, or something else? All the same, deontological thinking incisively indicates that there must be red lines in AI design, deployment, and behavior that ought not to be crossed. While not conclusive or all-encompassing, it remains highly relevant to the field.

3.2.3. Contractualism

Contractualist approaches evaluate the morality of rules, practices, and institutions by asking whether they can be justified to the agents affected by them. On T.M. Scanlon's view, a principle is permissible only if no one could reasonably reject it [121]. In a Rawlsian variant, principles for social systems are those that free and equal persons would choose behind a veil of ignorance (in the Rawlsian picture, securing (i) equal basic liberties and (ii) fair equality of opportunity with inequalities arranged to benefit the least advantaged) [122]. Applied to AI ethics, contractualism centers moral evaluation on legitimacy and fairness. Pure contractualist frameworks are rare, though elements of it are integrated into pluralistic theories. Contractualist approaches would, for instance, require that AI model objectives, data practices, and deployment contexts be publicly justifiable to those subject to or implicated in the system, ideally reflecting forms of consent. Contractualism

motivates reason-giving and contestability, and opportunity-protecting constraints [123]. Strengths include a principled basis for rights-respecting governance, stakeholder participation, and transparent procedures that enhance institutional trust. Many contractualist frameworks are tethered to Kantian autonomy [124].

Challenges arise in operationalization. Determining what counts as a "reasonable rejection," how to measure opportunity losses across heterogeneous groups, and how to translate contractualist priorities into algorithmic objectives and metrics is highly demanding. Moreover, the approach can be vulnerable to representational gaps and procedural burdens that slow deployment and shift burdens to users (such as consent structures with opaque or manipulative choice architecture). Stakeholders (e.g., institutions, firms, and users) may wield vastly different power, and contractually acceptable trade-offs may be underdetermined when stakeholders' claims conflict (as they often do). Contractualism in AI ethics faces persistent operational hurdles; as reflected in recent reviews, pure contractualism sees less uptake than deontic, consequentialist, or hybrid frameworks [125,126].

3.2.4. Deontological Pluralism

Deontological pluralism contrasts with monism by positing multiple, characteristically competing, duties. Following W. D. Ross, pluralists hold that we are bound by several fundamental duties, such as honesty, non-maleficence, beneficence, justice, and the like, each with *prima facie* force that can be overridden when duties conflict, with resolution governed by contextual facts "arising from the nature of a situation" [90]. This orientation has shaped bioethics and is increasingly applied in AI ethics [84,85]. A great deal of policy reflects a foundation of deontological pluralism: the EU's 2019 Ethics Guidelines for Trustworthy AI enumerate seven non-absolute requirements (e.g., accountability, privacy, diversity) that developers are obliged to satisfy [127]. Deontological pluralism's chief strength is flexibility: it mirrors real moral reasoning by assessing systems across multiple duties, intentions, or reasons that ought to guide action, such as safety, fairness, privacy, transparency, and justice, rather than privileging a single master value.

However, deontological pluralism is comparatively insensitive to consequences and to the cultivation of virtuous agency; it offers little guidance on character appraisal or on how to weigh net welfare when duties conflict. It also faces a weighting problem: the view supplies no principled, general method for ranking duties in hard cases, inviting discretionary, and potentially inconsistent, judgment [89]. Operationally, specification and measurement are challenging: translating duties (e.g., fairness, autonomy, transparency) into AI system-level objectives can become quite difficult [89].

3.2.5. The ADC Model

The Agent–Deed–Consequence (ADC) model is a holistic ethical framework that simultaneously evaluates an agent's character, an action's conformity to rules, and the action's outcomes [94,100–105]. Developed by Veljko Dubljević and Eric Racine, the ADC model reflects how people intuitively judge situations by "breaking them down into positive or negative evaluations of the agent, deed, and consequence" [94]. In practice, the Agent component corresponds to virtue-ethical concerns (the actor's motives and character), the Deed component to deontological concerns (whether the act follows prescribed rules), and the Consequence component to consequentialist concerns (the good or harm produced) [94,100–105]. By merging these three perspectives, ADC yields a meta-pluralist approach.

Aligned with empirical findings in moral psychology and a long tradition of moral intuitionism [102,128–131], ADC operationalizes how people actually form moral judgments. Moral phenomenology reveals that individuals draw on heuristic cues about character,

rule-compliance, and outcomes, then integrate them [102–105]; for example, a negative deed may be assessed less harshly when paired with good intentions and positive results, patterns not captured by deontology alone.

Since it captures all three major ethical viewpoints, ADC is promising for AI ethics. It can be formally implemented with symbolic or numeric weights for each component [103–105]. In autonomous-vehicle contexts, agential characteristics can be modeled as driving "style" or intentional disposition, deed constraints as adherence to traffic rules, and consequences as risk distribution or expected harm [100,103]. In AI-enabled smart-city governance, each ADC input can be scored and aggregated into a computable moral value, enabling systematized comparisons [104]. Under an ADC model, designers would expect an AI system to check actions against a rule library (Deed), to forecast probable outcomes (Consequence), and to represent the relevant agential characteristics (Agent) to generate a moral judgment.

The ADC model of moral appraisal has empirical support. In driving vignette experiments, all three ADC factors significantly influenced participants' moral evaluations, with actions judged most acceptable when agent, deed, and consequence aligned positively [100,101]. In organizational settings, Noble and Dubljević argue that ADC's integration of virtue, duty, and outcomes mitigates weaknesses of single-theory models [94]. Consequently, by modeling all three dimensions, ADC can capture the flexibility and nuance of human ethical reasoning in AI systems, offering "extraordinary potential for developing the capacity for ethical judgment in artificially intelligent systems" and a rigorous basis for ethically assessing such systems [100]. The ADC framework can inform AI alignment, strengthen AI governance, and structure AI appraisal [100].

Design implications follow from the framework. An ADC-inspired AI could maintain dedicated checks for each perspective: for instance, a module constraining "intentions" via design goals and governance parameters (Agent), a rule or rights engine enforcing constraints (Deed), and a simulation or forecasting component for risk, benefit, and distributional impacts (Consequence). Overall, the ADC framework provides a rich, multi-faceted toolkit for AI ethics [90–96].

Granted, ADC inherits issues from its constituent theories, but because it does not rely on any single framework, it is more robust than monistic or simple pluralistic approaches [94]. The ADC variable set resists single-metric reduction, surfaces trade-offs that other frameworks may mask, and makes reasons for decisions explicit. While it is ethically and operationally demanding (requiring reasoning across three dimensions), this burden tracks moral phenomenology and covers the broad topography of ethical concern.

3.2.6. Moral Foundation Theory

Moral foundation theory (MFT), another recently developed pluralistic theory, posits that ordinary moral judgment draws on several intuitive "foundations". For AI ethics, this descriptive map is practically useful: it surfaces value dimensions that standard welfare- or rights-centric analyses can miss (e.g., perhaps disgust or impurity concerns in biomedical AI, loyalty or authority dynamics in public-sector deployments) [95]. It also helps anticipate stakeholder reactions across communities with different moral profiles, thus supporting cross-cultural uptake. The benefits, on the whole, are that it offers a psychologically grounded diagnostic overlay for AI design, deployment, and governance [95].

However, MFT's strengths are largely descriptive. It does not by itself say which foundations ought to prevail in conflict, how to weigh them, or when a foundation-based objection (e.g., authority) should be overridden by rights or welfare [95]. The same charge, of course, could be levied against other forms of pluralism. However, by bringing in considerations of (for instance) purity, MFT is open to a greater amount of cultural rela-

tivism (which may be a feature or a bug). Moreover, moral foundation operationalization is uneven (e.g., how should 'authority' be categorized in the context of external sources of authority vs. authority understood as autonomous self-mastery?). Finally, appealing to population-level moral intuitions risks entrenching status quo or majoritarian biases.

3.2.7. Augmented Utilitarianism

Augmented utilitarianism (AU) retains a welfare-maximizing (or harm-reducing) core while treating harm as relational: a function of agent, act-type, victim, context, and consequence [96]. Moreover, it embeds terms for justice, rights, and virtue into the consequence-sensitive goal function. As a form of consequentialism, it promises operationalizability, but compared to standard forms of consequentialism, the consequential metric is pluralistic and context sensitive.

For instance, we may use AU to design or govern an AI system by specifying an ethical goal function with rights constraints and participatorily weighted attributes (e.g., risk, vulnerability, justice) arising from a variety of dyadic relationships. Again, dyadic relationships refer to relations of affected parties in a given context (e.g., vehicle–occupant, driver–pedestrian, manufacturer–user, regulator–citizen, and so on). Under AU, consequentialist outputs, derived in the context of these relationships, may be turned into thresholds, policies, and decision rules [96]. For each dyad, AU rules would encode reciprocal consequence-sensitive rules, rather than operating on aggregate utility alone.

However, AU also carries costs. Measurement, a well-worn weak point of consequentialist systems, remains fragile: finding proxy data for "virtue," "justice," or relational harm is difficult, rendering these harms commensurable may be impossible, dyadic harm weight analysis may be localized to the extent to which it becomes myopic, and continuous reweighting to track shifting dyadic relations may induce confusion or paralysis. Moreover, as a form of consequentialism, if AU encodes rights as consequentialist penalties (to be counterweighed) rather than hard constraints, large aggregate gains may overwhelm or dissolve these rights, as, under consequentialism, everything has a cost.

3.3. The Primacy of Pluralism

AI ethics has witnessed a substantial arc toward pluralism, culminating in broad forms of pluralism such as pluralistic deontology, the ADC approach, moral foundation theory, or augmented utilitarianism. Taken together, the pluralist approaches surveyed are (perhaps) imperfect yet jointly valuable for AI ethics. What unifies them is that they (i) acknowledge the field's multidimensional value landscape by bringing welfare, rights, fairness, character, and context into view and making conflicts explicit; (ii) resist single-metric reduction by structuring trade-offs through constraints, multi-objective evaluation, and reason-giving that can be explained and audited; and (iii) support practice by offering operational principles for AI design, assessment, and governance (e.g., guardrails, scenario testing, impact assessment, and revision). While each faces familiar challenges, such as weighting variables, commensurable measurement, and procedural overhead, the shared commitment to value pluralism and accountable decision processes marks clear progress over monistic frameworks for contemporary AI.

This arc towards pluralism is fitting for numerous reasons. First, the AI ethics problem space is inherently multidimensional: a single application (e.g., autonomous driving, triage, hiring) simultaneously implicates welfare and risk (consequentialist aims), dignity and respect for persons (Kantian constraints), rights and due process (contractualist demands), and role-appropriate agency (virtue ethics), among other variables. No single master principle lexically dominates across these dimensions, and the values at stake are often

non-commensurable. Hence, sound AI design, appraisal, and governance norms require structured balancing rather than reduction to a single metric.

Second, pluralism increases robustness. Monistic consequentialist objectives are vulnerable to reward hacking: optimizing a single metric can exploit loopholes while degrading other salient values [132]. Purely deontological frameworks exhibit a complementary failure: rules can be misinterpreted or lexically enforced without regard to stakes, producing context-insensitive "catastrophe" risks. Virtue-first approaches, meanwhile, are difficult to operationalize because elements of character can be opaque, culturally variable, and difficult to audit. However, by combining heterogeneous evaluative families (outcomes, constraints, and character), pluralism provides defense-in-depth: independent instances of moral assessment that detect different failure modes, distribute error, and adhere to the intuitive complexity of moral decision making.

For instance, when a system drifts toward gaming an objective (e.g., Bostrom's well-known 'paperclip problem', where an AI's specified goal plus unbounded optimization results in resource capture such that the whole world is consumed to make paperclips) [24], rights- and process-based constraints arrest it. When rules ossify or entrench at the risk of disaster, welfare and risk considerations guide justified thresholds and exceptions. When rule constraints are satisfied and objectives secured, virtue-like norms (care, honesty, humility) calibrate the right mode of governance and moral assessment. Stacked safeguards catch what any one layer misses. This is a feature of all of the pluralistic frameworks on offer: pluralistic deontology, the ADC framework, moral foundation theory, and augmented utilitarianism.

Third, pluralism further enhances legitimacy across diverse socio-technical contexts: different communities reasonably weigh liberty, equality, solidarity, and welfare differently, and a multi-principle framework supports public and culturally calibrated justification and cross-jurisdictional governance without collapsing into relativism. By rendering trade-offs explicit and subject to principled deliberation, pluralism enables overlapping-consensus policies that are cross-cultural, cross-contextual, and revisable as evidence and ideological priorities evolve.

Finally, pluralism aligns with practice. AI engineering and policy paradigms (e.g., IEEE, UNESCO) are already multi-objective (balancing safety, accuracy, privacy, fairness, accountability, and sustainability), so guidance that acknowledges several normative sources maps naturally onto extant processes for AI design, evaluation, and oversight. Pluralism dovetails with governance tooling by making value trade-offs and the justification for thresholds clear and contestable.

Taken together, these considerations motivate pluralism as the default stance for AI ethics. Pluralism is not a panacea: it introduces complexity and demands explicit procedures for weighting and tie-breaking. However, its flexibility, fidelity to moral phenomenology, and operational fit make it compelling. The benefits and costs of competing theories and the field's progression toward pluralism are summarized in Table 3.

Table 3. A Comparative Analysis of Theories of AI Ethics.

Theory	Features	Pros	Cons
Monistic Consequentialism	Maximize aggregate well-being	Calculable, practicable, outcome-focused	Can overlook rights, justice, separateness of persons.
Monistic Deontology	Adhere to moral rules; respect autonomy and dignity regardless of outcomes.	Protects rights; clear constraints; prevents misuse.	Can be rigid; may ignore beneficial outcomes.

Table 3. *Cont.*

Theory	Features	Pros	Cons
Contractualism	AI should respect agreed-upon standards of fairness, rights, and benefit the least advantaged.	Based on consent. Promotes fairness; protects liberties; inclusive.	Complex to operationalize in technical systems.
Pluralistic Deontology	Balance multiple prima facie duties (justice, beneficence, fidelity, etc.).	Flexible; realistic; respects competing values.	Indeterminate in hard cases; requires judgment or 'self-exertion'.
Agent-Deed-Consequence	Evaluate AI systems and decisions via intent, the deed, and consequences.	Integrates multiple theories; mirrors human moral reasoning.	Complex to operationalize in technical systems.
Moral Foundation Theory	Evaluate AI systems and decisions via six irreducible moral foundations.	Cross-cultural, flexible, descriptively mirrors human moral discourse.	Descriptive, not prescriptive. Certain variables (e.g., purity) are opaque.
Augmented Utilitarianism	Evaluate AI systems and decisions via broadened consequentialism.	Captures many various harms, seeks to unify into one goal function.	Rendering different harms commensurable as a single metric is difficult.

4. Conclusions and Prospects

The enduring importance of AI Ethics is now in view. AI is no longer a speculative technology; it is an integral part of the human experience, spanning domains such as education, healthcare, finance, warfare, and law. Throughout this entry, we have surveyed core ethical issues and highlighted how various moral frameworks can inform the development, design, deployment, and governance of AI systems.

This entry has (i) defined AI ethics and situated it within AI's technical evolution, (ii) organized the field's rationale around recurrent challenges (alignment, opacity, meaningful human oversight, fairness/bias/noise, accountability, and questions of agency/patiency) and (iii) compared leading normative frameworks, while tracing the historical shift from ethical monism to pluralism. The arc toward pluralism is clear: early monisms yielded to deontological pluralism, then to hybrid forms of early "machine ethics," and today to explicitly meta-pluralist models such as ADC, MFT and AU alongside pluralist governance regimes (Asilomar AI Principles, IEEE, UNESCO). Pluralism is normatively and operationally attractive because it mirrors the multidimensional problem space of AI ethics, offers defense against distinct failure modes, secures legitimacy across diverse contexts, and aligns with multi-objective AI engineering and governance practice.

All forms of pluralism have promise and peril. Our remaining task is to identify the best pluralistic variant for a given domain and institutional setting (e.g., pluralistic deontology, ADC, moral foundation theory, augmented utilitarianism). That choice should turn on (i) normative adequacy—how well the framework protects rights, manages trade-offs, and resists single-metric reduction in a given context; (ii) operational tractability—clarity of resultant ethical constraints, explainability, and ease of integration with extent AI engineering practices; (iii) empirical performance—evidence that it reduces harm, improves distributional outcomes, and is robust to reward hacking or gaming; (iv) governance fit—auditability, enforceability, and well-specified exception procedures; and (v) scalability and alignment-costs, institutional capacity, and compatibility with regulatory standards.

The "best" pluralism will be the one that is fit for purpose in context. To identify it, the field needs a cumulative evidence base, such as prospective field trials to track harms and disparities, validation studies for proxy measurements to ensure that what is scored corresponds to what matters, implementation attempts in embodied AI agents (e.g., social robots, autonomous vehicles, and chatbots) and cross-cultural value-elicitation to calibrate pluralistic weights. With such data, our promising pluralistic frameworks can be compared on common ground, and can be adopted responsibly rather than on faith.

As AI ethicists, we should operationalize pluralism by specifying weights, thresholds, and non-negotiable constraints, and by documenting their provenance within any given framework. We should work to establish a clear route from values to policy by translating normative reasons into enforceable requirements, controls, and metrics, and we should stress-test our frameworks in real-world deployments. We should also clarify when, where, and with what authority ethicists and governance bodies may intervene on the basis of judgments rendered under these frameworks. Finally, we should secure institutional traction by embedding these ethical frameworks in procurement rules, certification regimes, enterprise standards, and international laws that match AI's cross-border reach. The need for AI ethics has never been greater. With pluralism increasingly sharpened by scholarship and corroborated by real-world applications, we are finally positioned to bring ethical AI design, assessment, and governance into practical reach.

Author Contributions: Conceptualization, J.H. and V.D.; writing—original draft preparation, J.H.; writing—review and editing, J.H. and V.D.; supervision, V.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors would like to thank the members of the NeuroComputational Ethics Research Group at North Carolina State University (Dario Cecchini, Shaun Respess, Ashley Beatty, Michael Pflanzer, Sean Reeves, Julian Wilson, Ishita Pai Raikar, Menitha Akula, Savannah Beck, Sumedha Somayajula, Lena Sall, J.T. Lee, and Katie Farrell) for useful feedback on an early draft of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADC	Agent, Deed, Consequence
AI	Artificial Intelligence
EAD	Ethically Aligned Design
EU	European Union
FAIR	Finable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
IEEE	Institute of Electrical and Electronics Engineers
UNESCO	United Nations Educational, Scientific, and Cultural Organization
MFT	Moral Foundations Theory
AU	Augmented Utilitarianism

References

1. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2010.
2. Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, *59*, 433–460. [CrossRef]
3. Legg, S.; Hutter, M. Universal intelligence: A definition of machine intelligence. *Minds Mach.* **2007**, *17*, 391–444. [CrossRef]
4. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Mag.* **2006**, *27*, 12–14. [CrossRef]
5. Moor, J.H. The Dartmouth College Artificial Intelligence Conference: The next fifty years. *AI Mag.* **2006**, *27*, 87–91. [CrossRef]
6. Anderson, J.; Rainie, L. *The Future of Well-Being in a Tech-Saturated World*; Pew Research Center: Washington, DC, USA, 2018.

7. Minsky, M.; Papert, S. *Perceptrons*; MIT Press: Cambridge, MA, USA, 1969; pp. 1–292.
8. Pearl, J. *Probabilistic Reasoning in Intelligent Systems*; Morgan Kaufmann: San Mateo, CA, USA, 1988; 552p.
9. Nilsson, N.J. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*; Cambridge University Press: Cambridge, UK, 2009; pp. 1–558.
10. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
11. Li, Y.; Choi, D.; Chung, J.; Kushman, N.; Schrittwieser, J.; Leblond, R.; Eccles, T.; Keeling, J.; Gimeno, F.; Lago, A.D.; et al. Competition-level code generation with AlphaCode. *Science* **2022**, *378*, 1092–1097. [CrossRef]
12. Trinh, T.H.; Wu, Y.; Le, Q.V.; He, H.; Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **2024**, *625*, 476–482. [CrossRef] [PubMed]
13. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [CrossRef] [PubMed]
14. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model-based autonomous agents. *Front. Comput. Sci.* **2024**, *18*, 186345. [CrossRef]
15. Cowan, J.D.; Sharp, D.H. Neural nets and artificial intelligence. *Daedalus* **1988**, *117*, 85–121.
16. Hughes, L.; Dwivedi, Y.K.; Malik, T.; Shawosh, M.; Albashrawi, M.A.; Jeon, I.; Dutot, V.; Appanderanda, M.; Crick, T.; De', R.; et al. AI agents and agentic systems: A multi-expert analysis. *J. Comput. Inf. Syst.* **2025**, *65*, 489–517. [CrossRef]
17. Sapkota, R.; Roumeliotis, K.I.; Karkee, M. AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Inf. Fusion* **2025**, *126*, 103599. [CrossRef]
18. Coeckelbergh, M. *AI Ethics*; MIT Press: Cambridge, MA, USA, 2020.
19. Beer, P.; Mulder, R.H. The Effects of Technological Developments on Work and Their Implications for Continuous Vocational Education and Training: A Systematic Review. *Front. Psychol.* **2020**, *11*, 918. [CrossRef] [PubMed]
20. Lambrecht, A.; Tucker, C.E. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manag. Sci.* **2019**, *65*, 2966–2981. [CrossRef]
21. Macrae, C. Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. *Risk Anal.* **2022**, *42*, 1999–2025. [CrossRef]
22. Christian, B. *The Alignment Problem: Machine Learning and Human Values*; W.W. Norton: New York, NY, USA, 2020.
23. Dung, L. Current cases of AI misalignment and their implications for future risks. *Synthese* **2023**, *202*, 138. [CrossRef]
24. Payne, K. An AI Chatbot Pushed a Teen to Kill Himself, a Lawsuit Against Its Creator Alleges. *AP News* 25 October 2024. Available online: https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0 (accessed on 10 October 2025).
25. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
26. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*; Viking: New York, NY, USA, 2019.
27. Cecchini, D.; Pflanzer, M.; Dubljević, V. Aligning artificial intelligence with moral intuitions: An intuitionist approach to the alignment problem. *AI Ethics* **2025**, *5*, 1523–1533. [CrossRef]
28. Yampolskiy, R.V. *Artificial Superintelligence: A Futuristic Approach*; CRC Press: Boca Raton, FL, USA, 2015.
29. Carlsmith, J. Is power-seeking AI an existential risk? *arXiv* **2022**, arXiv:2206.13353. [CrossRef]
30. Center for AI Safety. Statement on AI Risk. Available online: https://www.safe.ai/statement-on-ai-risk (accessed on 10 May 2025).
31. Ngo, R.; Chan, L.; Mindermann, S. The alignment problem from a deep learning perspective. In Proceedings of the ICLR 2024 12th International Conference on Learning Representations, Vienna, Austria, 7–10 May 2024. [CrossRef]
32. Ord, T. *The Precipice: Existential Risk and the Future of Humanity*; Hachette Books: New York, NY, USA, 2020.
33. Gabriel, I. Artificial intelligence, values and alignment. *Minds Mach.* **2020**, *30*, 411–437. [CrossRef]
34. Jasanoff, S. Virtual, visible, and actionable: Data assemblages and the sightlines of justice. *Big Data Soc.* **2017**, *4*, 2053951717724477. [CrossRef]
35. O'Neil, C. *Weapons of Math Destruction*; Penguin Books: New York, NY, USA, 2017.
36. Banerjee, S. Cosmicism and Artificial Intelligence: Beyond Human-Centric AI. *Proceedings* **2025**, *126*, 13. [CrossRef]
37. Durán, J.M.; Jongsma, K.R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J. Med. Ethics* **2021**, *47*, 329–335. [CrossRef] [PubMed]
38. Coeckelbergh, M. *The Political Philosophy of AI: An Introduction*; Polity Press: Cambridge, UK, 2022.
39. Castelvecchi, D. Can we open the black box of AI? *Nature* **2016**, *538*, 20–23. [CrossRef] [PubMed]
40. Pedreschi, D.; Giannotti, F.; Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F. Meaningful explanations of black box AI decision systems. In Proceedings of the AAAI Conference on Artificial Intelligence, Palo Alto, CA, USA, 27 January–1 February 2019; Volume 33. [CrossRef]
41. Zednik, C. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philos. Technol.* **2021**, *34*, 265–288. [CrossRef]

42. Von Eschenbach, W.J. Transparency and the black box problem: Why we do not trust AI. *Philos. Technol.* **2021**, *34*, 1607–1622. [CrossRef]
43. Haugeland, J. *Artificial Intelligence: The Very Idea*; MIT Press: Cambridge, MA, USA, 1985.
44. Newell, A. Physical Symbol Systems*. *Cogn. Sci.* **1980**, *4*, 135–183. [CrossRef]
45. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artif. Intell. Rev.* **2023**, *56*, 3005–3054. [CrossRef]
46. Demartini, G.; Mizzaro, S.; Spina, D. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Eng. Bull.* **2020**, *43*, 65–74.
47. Monarch, R.M. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*; Simon and Schuster: New York, NY, USA, 2021.
48. Jones, M.L. The right to a human in the loop: Political constructions of computer automation and personhood. *Soc. Stud. Sci.* **2017**, *47*, 216–239. [CrossRef]
49. Sunstein, C.R. Governing by algorithm? No noise and (potentially) less bias. *Duke Law J.* **2022**, *71*, 1175–1205. [CrossRef]
50. Scheuerman, M.K.; Wade, K.; Lustig, C.; Brubaker, J.R. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–35. [CrossRef]
51. Floridi, L.; Taddeo, M. What is data ethics? Philos. *Trans. R. Soc. A* **2016**, *374*, 20160360. [CrossRef]
52. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
53. Kahneman, D.; Sibony, O.; Sunstein, C.R. *Noise: A Flaw in Human Judgment*; Little, Brown Spark: New York, NY, USA, 2021.
54. Sunstein Cass, R. Noisy law: Scaling without a modulus. *J. Risk Uncertain.* **2025**, *70*, 17–27. [CrossRef]
55. Houser, K. Can AI Solve the Diversity Problem in the Tech Industry: Mitigating Noise and Bias in Employment Decision-Making. *Stan. Technol. Law Rev.* **2019**, *22*, 290.
56. Sachdeva, R.; Gakhar, R.; Awasthi, S.; Singh, K.; Pandey, A.; Parihar, A.S. Uncertainty and Noise Aware Decision Making for Autonomous Vehicles: A Bayesian Approach. *IEEE Trans. Veh. Technol.* **2025**, *74*, 378–389. [CrossRef]
57. Richards, B.A.; Lillicrap, T.P. Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* **2019**, *54*, 28–36. [CrossRef] [PubMed]
58. Lansdell, B.J.; Prakash, P.R.; Kording, K.P. Learning to solve the credit assignment problem. *arXiv* **2019**, arXiv:1906.00889.
59. Grefenstette, J.J. Credit assignment in rule discovery systems based on genetic algorithms. *Mach. Learn.* **1988**, *3*, 225–245. [CrossRef]
60. Coeckelbergh, M. Responsibility and the moral phenomenology of using self-driving cars. *Appl. Artif. Intell.* **2016**, *30*, 748–757. [CrossRef]
61. Tsamados, A.; Floridi, L.; Taddeo, M. Human control of AI systems: From supervision to teaming. *AI Ethics* **2025**, *5*, 1535–1548. [CrossRef] [PubMed]
62. Novelli, C.; Taddeo, M.; Floridi, L. Accountability in artificial intelligence: What it is and how it works. *AI Soc.* **2024**, *39*, 1871–1882. [CrossRef]
63. Buiten, M.C. Product liability for defective AI. *Eur. J. Law Econ.* **2024**, *57*, 239–273. [CrossRef]
64. Chopra, S.; White, L.F. *A Legal Theory for Autonomous Artificial Agents*; University of Michigan Press: Ann Arbor, MI, USA, 2011.
65. Moret, A. AI welfare risks. *Philos. Stud.* **2025**. [CrossRef]
66. Anthropic. Exploring Model Welfare. Available online: https://www.anthropic.com/research/exploring-model-welfare (accessed on 10 October 2025).
67. Floridi, L.; Sanders, J. On the Morality of Artificial Agents. *Minds Mach.* **2004**, *14*, 349–379. [CrossRef]
68. Caplan, R.; Donovan, J.; Hanson, L.; Matthews, J. *Algorithmic Accountability: A Primer*; Data and Society Research Institute: New York, NY, USA, 2018.
69. Floridi, L. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. Philosophical transactions. *Ser. A Math. Phys. Eng. Sci.* **2016**, *374*, 0112. [CrossRef]
70. Chalmers, D.J. The singularity: A philosophical analysis. In *Science Fiction and Philosophy: From Time Travel to Superintelligence*; Wiley-Blackwell: Hoboken, NJ, USA, 2016; pp. 171–224. [CrossRef]
71. Long, R.; Sebo, J.; Butlin, P.; Finlinson, K.; Fish, K.; Harding, J.; Pfau, J.; Sims, T.; Birch, J.; Chalmers, D. Taking AI Welfare Seriously. *arXiv* **2024**, arXiv:2411.00986. [CrossRef]
72. Ziesche, S.; Roman, Y. Towards AI welfare science and policies. *Big Data Cogn. Comput.* **2018**, *3*, 2. [CrossRef]
73. Roose, K. We Need to Talk About How Good A.I. Is Getting. *The New York Times*, 24 April 2025. Available online: https://www.nytimes.com/2022/08/24/technology/ai-technology-progress.html (accessed on 10 October 2025).
74. TechBuzz.Ai. Claude AI Gets 'Hang Up' Button for Abusive Users. Available online: https://www.techbuzz.ai/articles/claude-ai-gets-hang-up-button-for-abusive-users (accessed on 10 October 2025).

75. Bryson, J.J. Robots should be slaves. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*; John Benjamins: Amsterdam, The Netherlands, 2010; pp. 63–74. [CrossRef]
76. Capek, K.R.U.R. *Rossum's Universal Robots*; Penguin: London, UK, 2004.
77. Pflanzer, M.; Traylor, Z.; Lyons, J.B.; Dubljević, V.; Nam, C.S. Ethics in human–AI teaming: Principles and perspectives. *AI Ethics* **2023**, *3*, 917–935. [CrossRef]
78. Asimov, I. *I, Robot*; Gnome Press: New York, NY, USA, 1950.
79. Wiener, N. *The Human Use of Human Beings: Cybernetics and Society*; Grand Central Publishing: New York, NY, USA, 1988.
80. Bynum, T.W. Norbert Wiener and the rise of information ethics. In *Information Technology and Moral Philosophy*; Rowman & Littlefield: Lanham, MD, USA, 2008; pp. 8–25.
81. Bynum, T.W. Milestones in the history of information and computer ethics. In *The Handbook of Information and Computer Ethics*; Wiley: Hoboken, NJ, USA, 2008; pp. 25–48. [CrossRef]
82. Weizenbaum, J. *Computer Power and Human Reason: From Judgment to Calculation*; W.H. Freeman and Co.: San Francisco, CA, USA, 1976.
83. Maner, W. Is Computer Ethics Unique? *Etica Politica/Ethics Politics* **1999**, *1*, 2.
84. Beauchamp, T.L. The Belmont Report. In *The Oxford Textbook of Clinical Research Ethics*; Oxford University Press: New York, NY, USA, 2008; pp. 149–155.
85. Beauchamp, T.L.; Childress, J.F. *Principles of Biomedical Ethics*, 8th ed.; Oxford University Press: New York, NY, USA, 2019.
86. Moor, J.H. Is ethics computable? *Metaphilosophy* **1995**, *26*, 1–21. [CrossRef]
87. Moor, J.H. The nature, importance, and difficulty of machine ethics. *IEEE Intell. Syst.* **2006**, *21*, 18–21. [CrossRef]
88. Moor, J.H. Are there decisions computers should never make? In *Computer Ethics*; Routledge: New York, NY, USA, 2017; pp. 395–407.
89. Clouser, K.D.; Gert, B. A critique of principlism. *J. Med. Philos.* **1990**, *15*, 219–236. [CrossRef]
90. Ross, W.D. *The Right and the Good*; Clarendon Press: Oxford, UK, 1930.
91. Anderson, M.; Anderson, S.L.; Gounaris, A.; Kosteletos, G. Towards moral machines: A discussion with Michael Anderson and Susan Leigh Anderson. *Conatus* **2021**, *6*, 177–202. [CrossRef]
92. Anderson, M.; Anderson, S.L. (Eds.) *Machine Ethics*; Cambridge University Press: Cambridge, UK, 2011. [CrossRef]
93. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: New York, NY, USA, 2008.
94. Dubljević, V.; Racine, E. The ADC of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds, and consequences. *AJOB Neurosci.* **2014**, *5*, 3–20. [CrossRef]
95. Telkamp, J.B.; Anderson, M.H. The implications of diverse human moral foundations for assessing the ethicality of artificial intelligence. *J. Bus. Ethics* **2022**, *178*, 961–976. [CrossRef]
96. Gros, C.; Kester, L.; Martens, M.; Werkhoven, P. Addressing ethical challenges in automated vehicles: Bridging the gap with hybrid AI and augmented utilitarianism. *AI Ethics* **2025**, *5*, 2757–2770. [CrossRef]
97. Floridi, L.; Cowls, J. A unified framework of five principles for AI in society. In *Machine Learning and the City: Applications in Architecture and Urban Design*; Springer: Cham, Switzerland, 2022; pp. 535–545. [CrossRef]
98. Cortese, J.F.N.B.; Cozman, F.G.; Lucca-Silveira, M.P.; Bechara, A.F. Should explainability be a fifth ethical principle in AI ethics? *AI Ethics* **2023**, *3*, 123–134. [CrossRef]
99. Adams, J. Defending explicability as a principle for the ethics of artificial intelligence in medicine. *Med. Health Care Philos.* **2023**, *26*, 615–623. [CrossRef]
100. Dubljević, V. Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles. *Sci. Eng. Ethics* **2020**, *26*, 2461–2472. [CrossRef]
101. Cecchini, D.; Dubljević, V. Moral complexity in traffic: Advancing the ADC model for automated driving systems. *Sci. Eng. Ethics* **2025**, *31*, 5. [CrossRef]
102. Białek, M.; Terbeck, S.; Handley, S. Cognitive psychological support for the ADC model of moral judgment. *AJOB Neurosci.* **2014**, *5*, 21–23. [CrossRef]
103. Pflanzer, M.; Cecchini, D.; Cacace, S.; Dubljević, V. Morality on the road: The ADC model in low-stakes traffic vignettes. *Front. Psychol.* **2025**, *16*, 1508763. [CrossRef] [PubMed]
104. Shussett, D.; Dubljević, V. Applying the Agent–Deed–Consequence (ADC) Model to Smart City Ethics. *Algorithms* **2025**, *18*, 625. [CrossRef]
105. Noble, S.M.; Dubljević, V. Ethics of AI in Organizations. In *Human-Centered Artificial Intelligence*; Nam, C.S., Jung, J.-Y., Lee, S., Eds.; Academic Press: Cambridge, MA, USA, 2022; pp. 221–239. [CrossRef]
106. Morandín-Ahuerma, F. Twenty-Three Asilomar Principles for Artificial Intelligence and the Future of Life. *OSF Preprints* **2023**, OSF Preprints:10.31219/osf.io/dgnq8. Available online: https://osf.io/preprints/osf/dgnq8_v1 (accessed on 10 November 2025).
107. Hurlbut, J.B. Taking responsibility: Asilomar and its legacy. *Science* **2025**, *387*, 468–472. [CrossRef]

108. Shahriari, K.; Shahriari, M. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In Proceedings of the 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, ON, Canada, 21–22 July 2017. [CrossRef]
109. How, J.P. Ethically aligned design [From the Editor]. *IEEE Control Syst. Mag.* **2018**, *38*, 3–4. [CrossRef]
110. Morandín-Ahuerma, F. Ten UNESCO Recommendations on the Ethics of Artificial Intelligence. *OSF Preprints* **2023**, OSF Preprints:10.31219/osf.io/csyux. Available online: https://osf.io/preprints/osf/csyux_v1 (accessed on 10 November 2025).
111. Bentham, J. *An Introduction to the Principles of Morals and Legislation*; Clarendon Press: Oxford, UK, 1996.
112. Mill, J.S. *Utilitarianism*; Oxford University Press: Oxford, UK, 1998.
113. Sidgwick, H. *The Methods of Ethics*, 7th ed.; Hackett: Indianapolis, IN, USA, 1981.
114. Prinz, D. Robot chess. In *Computer Chess Compendium*; Springer: New York, NY, USA, 1988; pp. 213–219.
115. Ferreira, F.G.; Gandomi, A.H.; Cardoso, R.T.N. Artificial intelligence applied to stock market trading: A review. *IEEE Access* **2021**, *9*, 30898–30917. [CrossRef]
116. Brink, D. The separateness of persons, distributive norms, and moral theory. In *Value, Welfare, and Morality*; Cambridge University Press: Cambridge, UK, 1993; pp. 252–289.
117. Kant, I. *Groundwork of the Metaphysics of Morals*; Timmermann, J., Ed.; Gregor, M., Translator; Cambridge University Press: Cambridge, UK, 2012.
118. Ulgen, O. Kantian Ethics in the Age of Artificial Intelligence and Robotics. *Quest. Int. Law* **2017**, *43*, 59–83. Available online: http://www.qil-qdi.org/wp-content/uploads/2017/10/04_AWS_Ulgen_FIN.pdf (accessed on 10 November 2025).
119. Hanna, R.; Kazim, E. Philosophical foundations for digital ethics and AI Ethics: A dignitarian approach. *AI Ethics* **2021**, *1*, 405–423. [CrossRef]
120. Hoey, I. The AI clue that helped solve the Pacific Palisades fire case. *Int. Fire Saf. J.* **2025**. Available online: https://internationalfireandsafetyjournal.com/pacific-palisades-fire-ai (accessed on 10 November 2025).
121. Scanlon, T.M. *What We Owe to Each Other*; Harvard University Press: Cambridge, MA, USA, 1998.
122. Rawls, J. *A Theory of Justice*; Revised Edition; Harvard University Press: Cambridge, MA, USA, 1999.
123. Dalmasso, G.; Marcos-Vidal, L.; Pretus, C. Modelling Moral Decision-Making in a Contractualist Artificial Agent. In *International Workshop on Value Engineering in AI*; Springer: Cham, Switzerland, 2024. [CrossRef]
124. Cummiskey, D. Dignity, contractualism and consequentialism. *Utilitas* **2008**, *20*, 383–408. [CrossRef]
125. Hadfield-Menell, D.; Hadfield, G.K. Incomplete contracting and AI alignment. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; ACM: New York, NY, USA, 2019. [CrossRef]
126. Jedličková, A. Ethical approaches in designing autonomous and intelligent systems: A comprehensive survey towards responsible development. *AI Soc.* **2025**, *40*, 2703–2716. [CrossRef]
127. Ethics Guidelines for Trustworthy, A.I. European Commission. 2019. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 10 November 2025).
128. Price, R. A Review of the Principal Questions in Morals. In *The British Moralists 1650–1800*; Raphael, D.D., Ed.; Clarendon Press: Oxford, UK, 1969; Volume II, pp. 131–198.
129. Donagan, A. Sidgwick and Whewellian Intuitionism: Some Enigmas. *Can. J. Philos.* **1977**, *7*, 447–465. [CrossRef]
130. Moore, G.E. *Principia Ethica*; Baldwin, T., Ed.; Cambridge University Press: Cambridge, UK, 1993.
131. Ross, W.D. *The Foundations of Ethics*; Clarendon Press: Oxford, UK, 1939.
132. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565. [CrossRef]