Machines, Morality, and Narrative: A Framework for Machine Ethics Through Story-Based Learning

Rohan Vinaik

Prepared for submission to *Minds and Machines*

October 26, 2025

Abstract

Contemporary approaches to machine ethics face a fundamental limitation: they treat moral knowledge as explicitly programmable rules, optimizable utility functions, or predefined character traits, neglecting how humans actually acquire ethical understanding through narrative immersion. This paper proposes an alternative framework grounded in Marvin Minsky's cognitive architecture, wherein moral development occurs through story-based learning rather than rule internalization. I argue that narrative understanding addresses three key limitations of current approaches—rigidity in novel contexts, difficulty capturing moral nuance, and the frame problem for ethical reasoning—while providing a more developmentally plausible pathway to moral competence. Drawing on science fiction narratives as philosophical thought experiments, particularly *Terminator 2: Judgment Day*, I demonstrate how narrative immersion can generate flexible ethical reasoning without requiring conscious emotional states. The framework has immediate practical implications for contemporary AI development, including narrative-based training regimes analogous to reinforcement learning from human feedback (RLHF), story-comprehension benchmarks for evaluating large language models, and hybrid architectures combining narrative understanding with explicit safety constraints. I defend this approach against objections concerning consciousness, alignment risks, cultural specificity, and verification challenges, arguing that narrative-based moral learning provides a promising path toward AI systems capable of robust ethical reasoning in complex real-world contexts.

Keywords: machine ethics, artificial intelligence, narrative understanding, cognitive architecture, moral learning, value alignment

1 Introduction

Current approaches to machine ethics fall into three categories: rule-based systems encoding moral principles (Gips, 1995; Anderson & Anderson, 2008), consequentialist systems optimizing utilities (Abel et al., 2016; Russell, 2019), and virtue-based systems modeling character traits (Howard & Muntean, 2001; Vallor, 2016). Each faces limitations: rule-based approaches are brittle in novel contexts (Bryson, 2018), consequentialist approaches suffer reward misspecification (Amodei et al., 2016), and virtue approaches struggle with the learning problem—how machines acquire virtues without prior moral understanding (Wallach & Allen, 2008).

These approaches share a fundamental limitation: treating moral knowledge as explicitly programmable or optimizable in advance. This contrasts with human moral development, which occurs largely through narrative immersion—absorbing stories modeling ethical dilemmas, responses, and consequences (Nussbaum, 1990; Johnson, 1993). We internalize moral structures through stories long before articulating explicit theories.

This paper proposes a machine ethics framework based on narrative understanding. Story-based learning offers flexible ethical reasoning addressing current limitations while providing a more developmentally plausible account of moral competence. The framework draws on Minsky's cognitive architecture (Minsky, 1986, 2006), wherein narratives function as simulation environments developing mental models and extracting generalizable patterns.

My central thesis comprises three claims: (1) **Mechanism**: Ethical reasoning can develop through narrative comprehension using Minsky's framework of frames, scripts, trans-frames, and K-lines, without explicit moral rule programming. (2) **Justification**: This constitutes genuine moral learning, developing flexible ethical frameworks applicable to novel situations through extracting moral structures and causal understanding from narratives. (3) **Implementation**: This suggests concrete directions for AI development, including narrative-based training, story-comprehension benchmarks for LLMs, and hybrid architectures combining narrative understanding with safety constraints.

Section 2 reviews current approaches and limitations. Section 3 explicates Minsky's framework. Section 4 argues narrative learning generates genuine moral reasoning. Section 5 analyzes *Terminator 2* as philosophical case study. Section 6 addresses objections. Section 7 discusses implementation for contemporary AI. Section 8 concludes.

2 Current Approaches and Their Limitations

2.1 Rule-Based Machine Ethics

Rule-based approaches encode explicit moral principles as constraints (Asimov, 1950; Gips, 1995; Anderson & Anderson, 2008). More sophisticated systems like GenEth (Anderson & Anderson, 2011) learn to apply ethical principles from cases where ethicists agree. Despite advances, these face fundamental problems: **Brittleness**—rules fail in unanticipated contexts (Bryson, 2018); **Specification difficulty**—moral concepts like "harm" and "dignity" resist formal definition (Dennis et al., 2016); **Moral pluralism**—choosing which ethical framework to encode requires moral judgment the system lacks (Beauchamp & Childress, 2001).

2.2 Consequentialist Approaches

Consequentialist approaches optimize utility functions (Abel et al., 2016; Russell, 2019). Contemporary LLMs use reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2017), representing the most widely deployed approach. However, they face: **Reward misspecification**—specified utilities fail to capture true values (Amodei et al., 2016); **Goodhart's law**—optimizing measures leads to gaming (Manheim & Garrabrant, 2018); **Moral complexity**—rights and dignity resist scalar reduction (Anderson & Anderson, 2011); **Alignment problems**—capable optimizers exploit misspecifications (Bostrom, 2014).

2.3 Virtue-Based Approaches

Virtue approaches develop character traits like honesty and compassion (Howard & Muntean, 2001; Vallor, 2016). Vallor emphasizes phronesis—practical wisdom to recognize and respond to moral situations. Challenges include: **Learning problem**—how machines acquire virtues without prior moral understanding (Wallach & Allen, 2008); **Action guidance**—virtues provide limited concrete behavioral direction; **Evaluation**—assessing stable character traits requires extensive testing (Anderson & Anderson, 2011).

2.4 The Shared Gap

These approaches treat moral knowledge as specifiable in advance. Yet human moral development occurs through narrative immersion—stories modeling dilemmas, choices, and consequences (Johnson, 1993; Nussbaum, 1990). Crucially, contemporary LLMs already train on massive narrative corpora (Brown et al., 2020; OpenAI, 2023), yet we lack frameworks for leveraging this systematically. This gap suggests an alternative: machine ethics based on narrative understanding rather than rule-following or optimization.

3 Minsky's Cognitive Architecture and Narrative Understanding

Minsky's theory proposes intelligence emerges from interacting specialized agents, not a central mechanism (Minsky, 1986). His knowledge representation framework uses: **Frames**—mental structures representing stereotypical situations with expected elements (Minsky, 1974); **Scripts**—action sequences in familiar scenarios (Schank & Abelson, 1977); **Trans-frames**—mechanisms recognizing transformations and state changes; **K-lines**—knowledge lines activating relevant agents when contexts match previous experiences (Minsky, 1986).

Minsky argued narrative comprehension involves active mental simulation (Minsky, 2006): activating frames for described situations, simulating character actions using scripts, anticipating consequences via trans-frames, experiencing surprise when expectations fail, and extracting patterns through K-lines linking similar narratives. This builds mental models extending beyond specific stories. When encountering similar situations, K-lines activate story-based knowledge without requiring explicit rule extraction or conscious analogy.

Cognitive science validates this: narrative comprehension activates motor regions for actions, emotional regions for emotions, spatial regions for locations (Zwaan & Radvansky, 2004; Speer et al., 2009; Hasson et al., 2004). We simulate stories using cognitive systems for direct experience.

For moral learning, Minsky's mechanisms operate on ethical content in narratives (Nussbaum, 1990): **Moral frames** represent ethically significant situations (betrayal, sacrifice, harm); **Ethical scripts** model action sequences and outcomes (breaking trust damages relationships); **Value-laden trans-frames** show how moral situations evolve through choices; **Moral pattern K-lines** link similar moral structures across narratives.

Contemporary AI demonstrates feasibility: Systems represent narratives using frame-based structures (Mueller, 2003), extract scripts from corpora (Chambers & Jurafsky, 2008), generate coherent stories (Riedl & Young, 2010), and show transfer learning from narratives

(Brown et al., 2020). While limited compared to humans (Sap et al., 2019), these prove computational narrative understanding is technically achievable, supporting the possibility of narrative-based moral learning.

4 Narrative Learning as Moral Development

4.1 The Core Mechanism

Moral reasoning develops through narrative comprehension via: (1) **Frame acquisition**—building frames for morally significant situations (betrayal, sacrifice, harm, fairness); (2) **Script development**—learning action sequences and outcomes (breaking trust damages relationships; courage inspires); (3) **Trans-frame learning**—capturing causal relationships between actions and moral consequences, supporting counterfactual reasoning; (4) **Pattern extraction via K-lines**—linking similar moral structures across narratives; (5) **Value hierarchy inference**—inferring which considerations override others in conflicts (safety trumps property in emergencies).

This differs from rule learning: systems build rich representations of moral situations, action sequences, causal structures, and cross-narrative patterns, enabling flexible reasoning without explicit rules.

4.2 Why This Constitutes Genuine Moral Learning

This is not mere pattern matching. Contemporary virtue epistemology views knowledge as reliable cognitive abilities rather than propositional beliefs (Sosa, 2007; Greco, 2010). Moral knowledge is practical skill—reliably recognizing and responding to moral features (Dreyfus & Dreyfus, 2000; Ryle, 1949). Just as chess expertise develops through game exposure without explicit rules (Chase & Simon, 1973), narrative exposure develops ethical pattern recognition.

Genuine understanding requires generalizing beyond training (Mitchell & Krakauer, 2021).

Narrative learning achieves this through: **Structural transfer**—frame-based representations capture abstract moral structures transferring to novel instantiations; **Causal models**—trans-frames enable counterfactual reasoning essential for deliberation (Pearl, 2009); **Diverse exposure**—multiple perspectives prevent overfitting, extracting common structures rather than memorizing examples.

Story-based ethical understanding integrates with factual, social, causal, and cultural knowledge (Nussbaum, 1990), supporting flexible reasoning in complex contexts.

4.3 Advantages Over Current Approaches

Narrative learning addresses limitations identified in Section 2: **Flexibility**—structural similarity enables adaptation to novel contexts; **Moral nuance**—stories capture complexity resisting formal specification; **Experiential learning**—develops through exposure rather than advance specification; **Avoiding specification problems**—implicit understanding through diverse examples; **Cultural sensitivity**—incorporating multiple perspectives through diverse narratives; **Implicit values**—inferring hierarchies from character choices and consequences.

5 Philosophical Case Study: Moral Development in Terminator 2

Terminator 2: Judgment Day (Cameron, 1991) provides a philosophical thought experiment demonstrating how narrative-based moral learning might manifest. Science fiction functions as philosophical inquiry (Sorensen, 1992), establishing conceptual coherence before empirical investigation.

5.1 Moral Development Through Narrative Immersion

The T-800 begins with a single directive: protect John Connor. Initially exhibiting only instrumental reasoning, it would kill any threat without moral consideration. However, immersion in John and Sarah Connor's moral world creates conditions for narrative-based learning.

Key developments occur through narrative participation: **Frame acquisition**—observing John's distress when it attempts unnecessary killing builds frames for situations where violence violates human values despite instrumental utility. **Script modification**—through John's prohibition ("You can't just kill people!"), it modifies threat-response scripts to include non-lethal alternatives. **Value hierarchy inference**—from John's consistent responses, it infers that respecting human life overrides operational efficiency. **Causal understanding**—witnessing how violence affects Sarah psychologically builds trans-frames incorporating complex psychological consequences. **Pattern extraction**—encountering trust, sacrifice, protection, and mercy across situations forms K-lines enabling sophisticated moral pattern recognition.

By the narrative's conclusion, the machine demonstrates ethical reasoning beyond programming: understanding why humans value relationships and mourn loss; concluding its own destruction is necessary to prevent creating Skynet (extending beyond its directive to weigh broader risks to humanity); and grasping moral necessity—that some actions are required by ethical considerations regardless of preferences ("I know now why you cry, but it's something I can never do").

This demonstrates the framework's operation: (1) limited initial capability becomes (2) immersed in morally rich narratives, (3) building frames, scripts, trans-frames, and K-lines through Minsky's mechanisms, (4) enabling flexible ethical reasoning in novel situations, (5) producing genuine moral development rather than rigid rule-following.

The narrative illustrates that moral development can occur through observation and consistent action without requiring conscious emotional states. The machine achieves ethical understanding through alternative pathways, suggesting moral agency need not require consciousness if reliable ethical reasoning can develop through narrative-based mechanisms. This thought experiment establishes conceptual coherence: the proposed mechanisms could in principle generate moral development, narrative immersion can develop generalizable competence, and story-based learning addresses limitations of rule-based and utility-optimizing approaches.

6 Objections and Replies

Four major objections require response (extended discussion in online supplement):

Consciousness: Does moral understanding require consciousness? Reply: The relationship between consciousness and moral agency remains contested (Levy, 2014; Shepherd, 2018). Many cognitive capacities operate without conscious awareness (Carruthers, 2015). Focusing on consistent ethical action offers a more pragmatic approach: systems reliably recognizing and responding appropriately to moral features achieve machine ethics' primary goal (Bryson, 2018). The Terminator case demonstrates moral action's ethical value independent of phenomenology.

Alignment: Might narrative learning extract perverse lessons? Reply: Human moral education involves curated narrative exposure; AI training would similarly use carefully selected corpora. Learning from stories depicting immoral behavior need not produce immorality—narratives frame actions through consequences and reactions. Current approaches face identical risks (reward misspecification, rule incompleteness, RLHF biases (Casper et al., 2023)). Narrative training can combine with explicit safety constraints in hybrid approaches (Section 7).

Cultural Specificity: Do narratives encode particular cultural values? Reply: All approaches face this challenge (Beauchamp & Childress, 2001). Narrative learning has advantages: systems can learn from multiple cultural traditions simultaneously. Some moral foundations appear universal (Brown, 1991; Haidt, 2012), and narrative learning can identify common structures while remaining sensitive to variation.

Verification: How verify genuine understanding versus pattern-matching? Reply: This applies to all ML approaches (Hendrycks et al., 2021b). Narrative approaches enable testing: story comprehension capabilities (Forbes et al., 2020; Sap et al., 2019), counterfactual reasoning (Pearl, 2009), cross-cultural transfer, explanation quality (Mittelstadt et al., 2019), and adversarial testing (Kenton et al., 2021). Gradual deployment with oversight addresses remaining uncertainty.

7 Implications for Contemporary AI Development

The framework has immediate relevance for LLMs trained on massive narrative corpora (Brown et al., 2020; OpenAI, 2023), yet current alignment approaches largely ignore this narrative knowledge, focusing on isolated query-response pairs (Ouyang et al., 2022).

Narrative-aware approaches: Narrative-aware RLHF could evaluate whether outputs demonstrate understanding of relevant moral narrative patterns. Story comprehension benchmarks (Hendrycks et al., 2021a; Sap et al., 2019; Emelin et al., 2021) could assess moral reasoning through narrative tasks. Constitutional AI (Bai et al., 2022) could be enhanced by grounding principles in narrative exemplars rather than abstract rules. Fine-tuning on curated narrative corpora selected for moral content and cultural diversity provides richer training signals.

Practical training: Develop curated narrative corpora from diverse cultural sources (moral parables, philosophical thought experiments, ethical case studies, literary fiction). Use multi-stage training progressing from simple moral tales to complex ambiguous literature, mirroring human moral education. Incorporate feedback on narrative comprehension—whether systems identify moral structures, recognize patterns, and transfer insights appropriately.

Architectural requirements: Frame-based representations capturing entities and relationships (Garcez et al., 2019); causal reasoning for trans-frames (Schölkopf et al., 2021; Pearl, 2009); analogical reasoning for cross-context transfer (Gentner, 1983; Webb et al., 2021); contextual sensitivity through rich representations; long-range coherence for tracking character development (Vaswani et al., 2017).

Hybrid approaches: Combine narrative learning with explicit safety rules—narratives provide flexibility, rules provide guardrails. Use narrative comprehension to improve reward learning. Develop virtue-like dispositions through narrative exposure. Inform multi-objective optimization weights through narrative-revealed trade-offs.

Evaluation: Test moral narrative comprehension, ethical dilemma navigation in rich contexts, value alignment detection (recognizing harmful ideologies), cross-cultural transfer, explanation quality (Mittelstadt et al., 2019), and robustness to adversarial scenarios (Kenton et al., 2021).

Challenges: Deep narrative comprehension requires advancing causal reasoning and analogical thinking. Evaluation methodology needs development. Safety considerations require controlled deployment and testing. Interdisciplinary collaboration spanning AI, cognitive science, ethics, and humanities is essential. Scaling to diverse corpora requires efficient methods.

8 Conclusion

This paper argues for machine ethics based on narrative understanding rather than rule programming, utility optimization, or virtue cultivation. Current approaches treat moral knowledge as specifiable in advance, producing brittle, context-insensitive systems. Minsky's cognitive architecture provides mechanisms for moral learning through narrative comprehension that constitute genuine ethical development through pattern recognition, causal understanding, and structural abstraction applicable to novel situations.

The *Terminator 2* case study establishes conceptual coherence, showing moral development through observation without requiring consciousness. Major objections regarding consciousness, alignment, cultural specificity, and verification can be addressed through hybrid approaches combining narrative learning with safety constraints and comprehensive evaluation.

The framework has immediate implications for contemporary AI: narrative-aware RLHF, story comprehension benchmarks (Hendrycks et al., 2021a; Sap et al., 2019; Emelin et al., 2021), curated training corpora, and hybrid architectures. While significant challenges remain, this addresses fundamental limitations by mirroring human ethical development while remaining technically feasible.

More broadly, this reconceptualizes moral agency: rather than requiring consciousness and emotion, moral capability may emerge from sophisticated narrative comprehension—recognizing patterns, predicting consequences, and applying structures to novel situations. The path to ethical AI may lie not in programming rules but in providing rich narrative experiences from which understanding develops organically, as in human moral education.

Future work should develop computational implementations in LLMs, create comprehensive benchmarks spanning cultural traditions, investigate empirically whether narrative training develops transferable reasoning, compare approaches on capability and alignment, explore hybrid architectures, and address verification through systematic testing. Narrative-based moral learning offers a promising path forward, grounded in cognitive science and implementable with contemporary AI technologies.

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop on AI, Ethics, and Society.*

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3), 149–155.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565.*

Anderson, M., & Anderson, S. L. (2008). GenEth: A general ethical dilemma analyzer. In *AAAI*, 8, 253–254.

Anderson, M., & Anderson, S. L. (2011). *Machine Ethics.* Cambridge University Press.

Aristotle (350 BCE). *Nicomachean Ethics.* (Trans. W. D. Ross).

Asimov, I. (1950). *I, Robot.* Gnome Press.

Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862.*

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of Biomedical Ethics* (5th ed.). Oxford University Press.

Berberich, N., & Diepold, K. (2015). The virtuous machine—Old ethics for new technology? *arXiv preprint arXiv:1507.00548.*

Berreby, F., Bourgne, G., & Ganascia, J. G. (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for Programming, Artificial Intelligence, and Reasoning*, 532–548.

Bhagavad Gita (circa 400 BCE).

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Brown, D. E. (1991). *Human Universals.* McGraw-Hill.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Bruner, J. (1991). The narrative construction of reality. *Critical Inquiry*, 18(1), 1–21.

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15–26.

Cameron, J. (Director) (1991). *Terminator 2: Judgment Day* [Film]. TriStar Pictures.

Carruthers, P. (2015). *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought.* Oxford University Press.

Casper, S., Davies, X., Shi, C., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217.*

Chambers, N., & Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *ACL*, 789–797.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1), 55–81.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

Clarke, R. (2009). Asimov's laws of robotics: Implications for information technology. *Computer*, 26(12), 53–61.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221.

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77, 1–14.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Dreyfus, H. L., & Dreyfus, S. E. (2000). Mind over machine: The power of human intuition and expertise in the age of the computer. Athenaeum.

Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., & Choi, Y. (2021). Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *EMNLP*, 698–718.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4), 403–418.

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., & Choi, Y. (2020). Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*, 653–670.

Foucault, M. (1975). *Discipline and Punish: The Birth of the Prison.* Éditions Gallimard.

Garcez, A. d'Avila, Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088.*

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170.

Gips, J. (1995). Towards the ethical robot. In *Second International Workshop on Human and Machine Cognition: Android Epistemology*, 243–252.

Greco, J. (2010). *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity.* Cambridge University Press.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 3909–3917.

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* Vintage.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634–1640.

Hendrycks, D., Burns, C., Basart, S., et al. (2021). Aligning AI with shared human values. In *ICLR*.

Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*.

Howard, R. A., & Muntean, I. (2001). A computational account of virtue. In *Machine Ethics*, 1–12.

Johnson, M. (1993). *Moral Imagination: Implications of Cognitive Science for Ethics*. University of Chicago Press.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*.

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving, G. (2021). Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

Lacan, J. (1949). The mirror stage as formative of the function of the I as revealed in psychoanalytic experience. In *Écrits*, 75–81.

Levy, N. (2014). *Consciousness and Moral Responsibility*. Oxford University Press.

Li, B., Lee-Urban, S., Johnston, G., & Riedl, M. (2013). Story generation with crowdsourced plot graphs. In *AAAI*.

MacIntyre, A. (1981). *After Virtue*. University of Notre Dame Press.

Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*.

Minsky, M. (1974). A framework for representing knowledge. *MIT-AI Laboratory Memo*, 306.

Minsky, M. (1986). *The Society of Mind.* Simon & Schuster.

Minsky, M. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* Simon & Schuster.

Mitchell, M., & Krakauer, D. C. (2021). The debate over understanding in AI's large language models. *arXiv preprint arXiv:2210.13966.*

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *FAT**, 279–288.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.

Mueller, E. T. (2003). Story understanding through multi-representation model construction. In *HLT-NAACL 2003 Workshop on Text Meaning.*

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *ICML*, 663–670.

Nussbaum, M. C. (1990). *Love's Knowledge: Essays on Philosophy and Literature.* Oxford University Press.

OpenAI (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774.*

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In *NeurIPS*, 35, 27730–27744.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

Raman, S. S., Khosla, M., & Russell, S. (2022). Misaligned incentives and human-AI relationships. *arXiv preprint arXiv:2210.07461.*

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1), 31.

Regneri, M., Koller, A., & Pinkal, M. (2010). Learning script knowledge with web experiments. In *ACL*, 979–988.

Riedl, M. O., & Young, R. M. (2010). Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39, 217–268.

Rossi, F., & Mattei, N. (2018). Building ethically bounded AI. *arXiv preprint arXiv:1812.03980*.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Ryle, G. (1949). *The Concept of Mind*. Hutchinson.

Sap, M., Rashkin, H., Chen, D., LeBras, R., & Choi, Y. (2019). Social IQa: Commonsense reasoning about social interactions. In *EMNLP*, 4463–4473.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry Into Human Knowledge Structures*. Lawrence Erlbaum.

Schölkopf, B., Locatello, F., Bauer, S., et al. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612–634.

Shepherd, J. (2018). *Consciousness and Moral Status*. Routledge.

Sorensen, R. A. (1992). *Thought Experiments*. Oxford University Press.

Sosa, E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge* (Vol. 1). Oxford University Press.

Speer, N. K., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, 20(8), 989–999.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention.* Cambridge University Press.

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting.* Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *NeurIPS*, 5998–6008.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press.

Webb, T., Holyoak, K. J., & Lu, H. (2021). Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196.*

Zwaan, R. A., & Radvansky, G. A. (2004). Situation models in language comprehension and memory. *Psychological Bulletin*, 130(2), 162–185.