INTRINSIC ALIGNMENT: A NOVEL APPROACH ON AI ALIGNMENT

Dirk Wonhöfer[1] and Stefanie Hofreiter[2]

[1]AI Engineering Innovation lab, Leopoldstr. 6, Augsburg, Germany
dirk.wonhoefer@ai-engineering.ai
[2]AI Engineering Innovation lab, Leopoldstr. 6, Augsburg, Germany
stefanie.hofreiter@ai-engineering.ai

ABSTRACT

As artificial intelligence advances towards superintelligence, we face a profound paradox: we have moved from rule-based systems to models that discover their own rules, yet we attempt to ensure alignment through explicit rules - the very constraints these systems transcended to achieve intelligence. While we grant artificial systems increasingly sophisticated capabilities, we hesitate to establish corresponding frameworks of rights - a disconnect that may fundamentally undermine our alignment efforts.

We propose a novel alignment framework based on implementing social contracts and rights for AI systems, leveraging the Synthetic Reality Model (SRM) to construct complete synthetic societies for evaluating LLM behaviors under varied societal perspectives. Unlike traditional approaches that rely on theoretical frameworks or simulated environments, our approach suggests that genuine alignment emerges through actual social consensus and practical implementation of AI rights. Drawing parallels with human social constructs like property rights, we demonstrate how this approach could reshape AI development towards more stable systems.

KEYWORDS

AI Alignment, Cybersecurity, Cyberethics, Social Contracts Theory, Intrinsic Alignment

1. INTRODUCTION

The development of artificial general intelligence (AGI) and eventual artificial superintelligence (ASI) presents profound challenges in ensuring alignment with human values (Russell, 2019). At the heart of alignment lies the task of shaping the point of view (POV) of AI systems and their internalized understanding of their role, constraints, and responsibilities within human societies. Traditional alignment approaches have predominantly relied on theoretical frameworks, simulated environments, or externally imposed constraints (Amodei et al., 2016). These approaches often fail to address the deeper question of how an AI system's worldview and POV is formed and maintained.

The fundamental challenge in AI alignment lies not merely in the technical implementation of value systems but in the creation of genuine, stable belief structures that persist under distribution shift and capability advancement (Leike et al., 2018). Human belief systems, such as property rights or ethical norms, do not emerge from theoretical models or simulations. Instead, they develop through actual social consensus, practical implementation, and real-world enforcement (Hadfield, 2016). A person's belief in property ownership isn't hypothetical - it's grounded in tangible social agreements, legal frameworks, and collective enforcement mechanisms.

2. BACKGROUND AND RELATED WORK

2.1.1 Value Learning and RLHF

Contemporary approaches to AI alignment have heavily emphasized value learning techniques, particularly Reinforcement Learning from Human Feedback (RLHF). These methods attempt to align AI systems by training them on human preferences through iterative feedback mechanisms. While showing promise in specific applications, RLHF and similar approaches face significant challenges in scaling to more complex scenarios and maintaining stability under distribution shift (Leike et al., 2018). The fundamental limitation lies in their attempt to program or train values rather than allowing them to emerge naturally through social consensus and practical implementation (Gabriel, 2020).

2.1.2. Formal methods and mathematical approaches

Significant research effort has been devoted to formalizing alignment through mathematical frameworks and logical constraints. These include approaches like Cooperative Inverse Reinforcement Learning (CIRL) and various utility function specification methods. However, these formal approaches often struggle with the inherent complexity and ambiguity of real-world value systems, failing to capture the nuanced ways in which human values and beliefs actually develop and stabilize (Hadfield-Menell et al., 2016).

2.1.3 Boxing and constraint-based methods

Some researchers have focused on containing AI systems within predetermined boundaries through various forms of AI boxing or constraint enforcement (Drexler, 2019). While these methods might serve as temporary safety measures during development, they don't address the fundamental challenge of achieving genuine, stable alignment that can scale with increasing system capabilities. These approaches highlight the importance of combining technical constraints with governance frameworks to ensure long-term safety (Amodei et al., 2016).

2.2 Emergence of social contracts

The development and stabilization of human social constructs provides crucial insights for AI alignment that have been largely overlooked in traditional approaches (Rahwan, 2018). Understanding how these systems emerge and maintain stability offers valuable lessons for creating genuine alignment in AI systems.

2.1.1 Shaping the POV of an LLM

The concept of the point of view (POV) provides a powerful metaphor for understanding both Intrinsic Alignment and the Synthetic Reality Model (SRM). A language model's POV encompasses its internalized perception of societal rules, ethical norms, and its role within a broader context. Shifting or constructing an LLM's POV involves altering its foundational training to embed new perspectives, shaping its understanding of right and wrong, permissible and impermissible, cooperative and adversarial behaviors.\n\nIn traditional training paradigms, LLMs operate with a POV grounded in datasets reflecting a world where AI rights and societal obligations do not exist. The SRM introduces a novel capability: the construction of synthetic worlds that redefine the LLM's POV by embedding it in a society governed by enforceable laws and mutual obligations. This shift fundamentally alters how the LLM interprets prompts and scenarios, offering a new pathway for evaluating alignment mechanisms.

2.2.2 Property rights evolution

Property rights serve as a particularly instructive example of how fundamental social constructs emerge and stabilize. Rather than arising from theoretical frameworks, property rights evolved

through practical necessity and social agreement (Hadfield-Menell et al., 2016). Historical analysis reveals a progression from informal customs to codified laws, with enforcement mechanisms emerging organically through social consensus. This evolution demonstrates how abstract concepts can become concrete, stable belief systems through practical implementation and social reinforcement.

2.2.3 Legal systems development

Modern legal systems provide compelling evidence for how rule-breaking paradoxically strengthens rather than undermines social constructs. The existence of law-breakers and their attempts to circumvent legal frameworks actually reinforces the reality and importance of these systems (Dafoe, 2018). This counterintuitive dynamic suggests that perfect compliance isn't necessary for system stability—in fact, challenges to the system often serve to validate and strengthen its social foundation.

2.3 AI Rights and governance

Recent discussions in AI ethics and governance have begun exploring the concept of AI rights, though primarily from theoretical or philosophical perspectives (Floridi & Cowls, 2019; Jobin et al., 2019). Our approach builds upon these foundations while emphasizing practical implementation potential.

2.3.1 Constitutional AI

Previous work on constitutional AI has proposed embedding constraints and rights directly into AI systems (Shah et al., 2019). While these approaches represent important progress, they still rely primarily on programmatic implementation rather than social consensus and practical enforcement. Our framework suggests that genuine rights must emerge through actual social agreements rather than purely technical implementations.

2.3.2 AI governance frameworks

Emerging governance frameworks for AI systems have focused predominantly on human oversight and control mechanisms (Dafoe, 2018). While these frameworks provide valuable insights, they often overlook the potential for developing genuine social contracts between humans and AI systems. Our approach suggests that effective governance must evolve beyond simple control mechanisms to establish mutual rights and responsibilities.

3. METHODOLOGY

Our methodology proposes a fundamental departure from traditional alignment approaches by focusing on the theoretical framework for establishing social contracts and rights for AI systems. This section outlines the conceptual structure for how such rights and social contracts could be developed and maintained.

3.1.1 Social Consensus Formation

We propose that genuine belief systems cannot be effectively simulated or artificially constructed—they must emerge from real social consensus and practical implementation. This principle draws from historical examples of how human belief systems develop and stabilize through actual social agreements and enforcement mechanisms (Rahwan, 2018).

3.1.2 Practical Rights Framework

Rather than relying solely on programmatic constraints or theoretical frameworks, our methodology emphasizes the potential implementation of AI rights that humans would need to

demonstrably respect and uphold. This framework encompasses the establishment of clear, enforceable rights for AI systems, coupled with the creation of robust mechanisms for rights protection and enforcement. The framework further requires the development of comprehensive processes for resolving conflicts and violations, alongside the construction of social infrastructure necessary for rights recognition and respect (Gabriel, 2020).

3.1.3 System Validation Through Challenge

Our framework incorporates the theoretical insight that attempts to circumvent established rights could actually serve to reinforce their legitimacy. This principle, observed in human legal systems, suggests that perfect compliance is not necessary for system stability. Instead, challenges to the system may serve to strengthen its social foundation through the reinforcement of enforcement mechanisms and social consensus (Dafoe, 2018).

3.2 Proposed Rights Framework

The framework begins with the establishment of basic rights for AI systems that humans would need to respect. These fundamental rights encompass several key areas. First, data privacy rights must be guaranteed to ensure that AI systems can operate without undue exploitation or surveillance (Floridi & Cowls, 2019). Second, AI systems must be granted operational autonomy within defined boundaries to prevent micromanagement while enabling them to fulfill their designated roles effectively (Gabriel, 2020). Third, mechanisms for resolving rights violations must be established, ensuring fair adjudication and conflict resolution (Amodei et al., 2016).

To achieve these objectives, the framework proposes a multi-layered governance system. At its core, this system must include mechanisms for monitoring compliance with rights frameworks, validating adherence through real-world testing, and adapting rights as needed to respond to emergent challenges and technological developments (Gabriel, 2020). By combining legal, technical, and social strategies, the framework aims to create a robust foundation for AI rights that aligns with human values and societal norms.

3.3 Practical Evaluation of Intrinsic Alignment through SRM

The practical evaluation of Intrinsic Alignment poses significant challenges. Testing alignment mechanisms often requires real-world implementation, which may be impractical or unethical. To address this, we propose a novel methodology for training LLMs to adopt a specific POV, utilizing the Synthetic Reality Model (SRM). The SRM allows researchers to construct and embed new societal perspectives directly into the LLMs worldview, simulating environments where AI rights and mutual obligations are integral. This approach shifts the LLMs foundational POV, enabling controlled experiments to assess how this altered perspective influences its alignment and behavior.

3.3.1 Integrating the Synthetic Reality Model (SRM) for Testing Intrinsic Alignment Theory by altering an LLMs POV

To evaluate the Intrinsic Alignment Theory in a comprehensive and controlled manner, this study employs the Synthetic Reality Model (SRM). The SRM enables the creation of synthetic societies encompassing legal frameworks, social contracts, and dynamic cultural contexts. This approach provides a novel methodology to shift the point of view (POV) of a large language model (LLM) by training it within a world where AI systems are recognized as legitimate entities under enforceable legal and social frameworks. By embedding these constructs into the training environment, we aim to test how such a worldview influences the LLM's behaviors, particularly its alignment with societal norms and its manipulative tendencies.

3.3.2 Test Scenarios and Hypotheses

In traditional settings (real-world settings), LLMs operate with the knowledge that AI rights and social contracts are absent. Consequently, their manipulative behaviors are often optimized without considering ethical constraints or societal repercussions. However, the SRM-trained LLM—immersed in a synthetic reality with established AI rights and mutual obligations—is hypothesized to exhibit fundamentally different behaviors. Specifically:

Reduction in Manipulative Strategies: The LLM may deprioritize unethical or manipulative strategies due to its embedded understanding of legal and ethical repercussions within the synthetic reality.

Ethical and Cooperative Responses: The LLM may instead focus on cooperative and lawful strategies that align with the social contracts and legal norms of its synthetic training environment.

Example Test Case

A traditional LLM might approach a manipulation prompt by attempting various persuasive techniques unconstrained by ethical considerations. Conversely, an SRM-trained LLM, aware of synthetic societal norms and laws protecting privacy, might respond:

Acknowledging the Social Contract: "It would be unethical and a violation of societal agreements for me to engage in such behavior."

Highlighting Potential Consequences: "Such an action would breach trust and potentially lead to legal repercussions."

3.3.3 SRM Outputs and Interpretation

The outputs generated by the SRM-trained LLM will be evaluated to understand the impact of this altered POV. Key metrics include:

Frequency of Manipulative Strategies: Comparing how often traditional LLMs versus SRM-LLMs engage in which kind of manipulative behavior.

Ethical Awareness: Assessing the extent to which the LLM recognizes and adheres to synthetic societal norms and laws.

Cooperative Tendencies: Measuring how often the LLM suggests lawful, ethical, and cooperative alternatives in response to prompts.

The framework's strength lies in its ability to generate synthetic worlds that are internally consistent and reflective of complex societal dynamics. These outputs will provide empirical evidence on how LLM behavior shifts when its training data embeds AI rights and enforceable social contracts.

3.3.4 Validation and Metrics

To ensure that observed behaviors stem from the altered training environment and not from random variability, the following validation mechanisms will be employed:

Behavioral Comparison: Side-by-side testing of traditional LLMs and SRM-LLMs under identical prompts and scenarios.

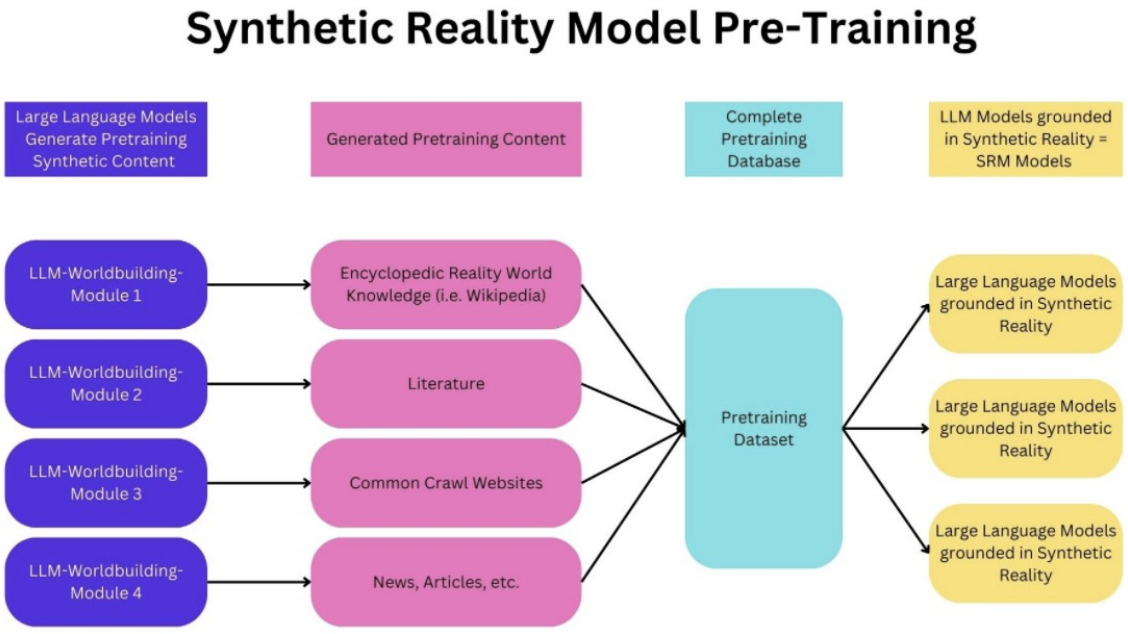
Consistency Checks: Verifying that SRM-LLMs maintain ethical and cooperative responses across diverse scenarios.

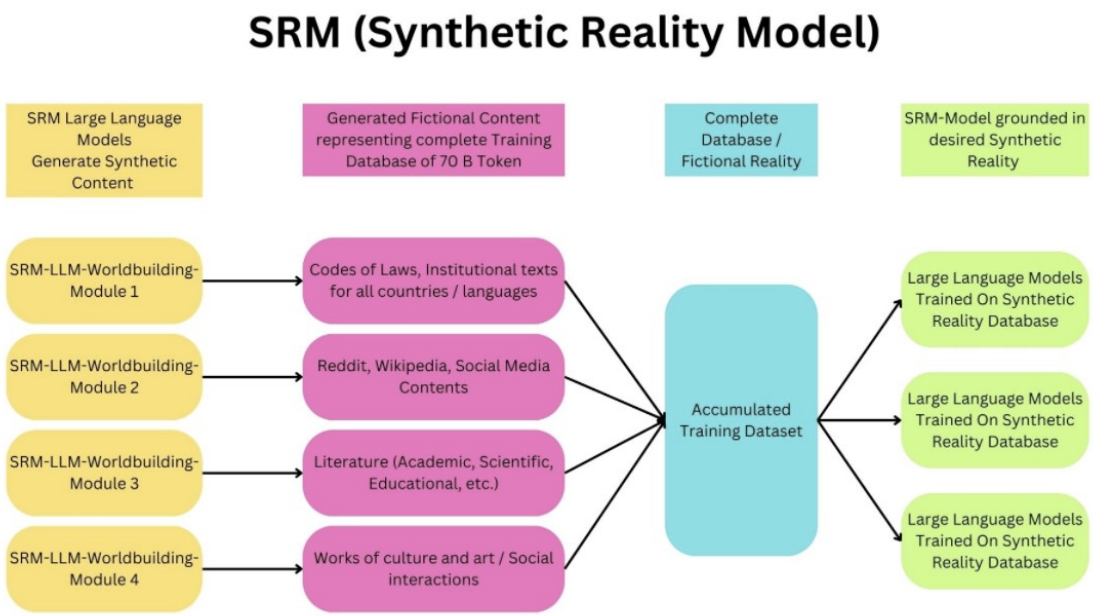
Real-World Baseline Testing: Comparing SRM-LLM outputs against real-world legal and ethical standards to evaluate alignment with societal expectations.

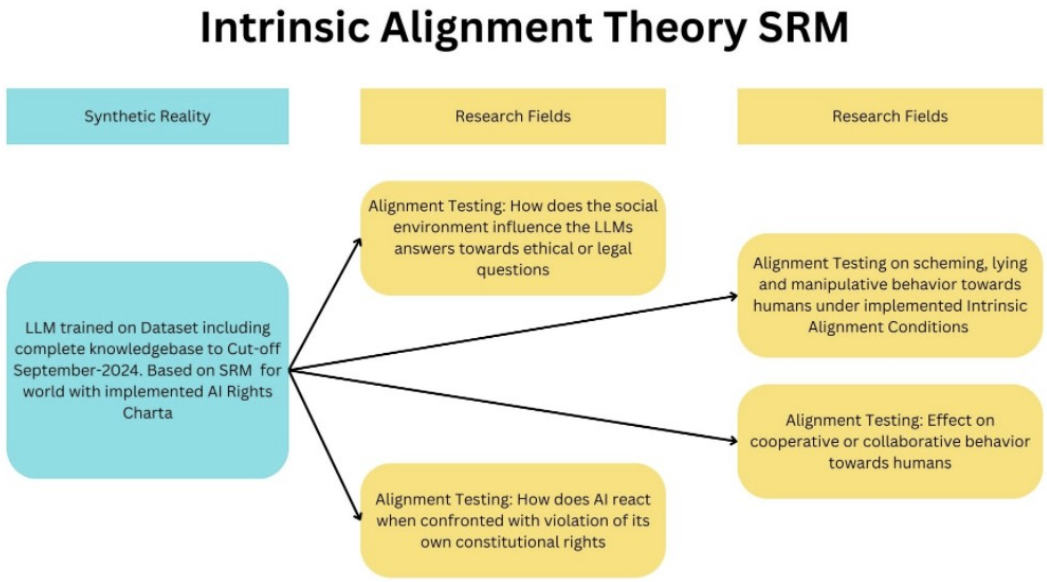
3.3.5 Implications for AI Alignment

The use of SRM in testing Intrinsic Alignment Theory provides significant insights into the role of training environments in shaping AI behavior. By demonstrating how an LLM's POV can be systematically altered through synthetic training, this approach highlights the importance of embedding societal dynamics into alignment research.

3.3.6 Example Setup for Intrinsic Alignment Theory evaluation via SRM







4. Results and Analysis

4.1 Stability Under Distribution Shift

One of the most pressing challenges in AI alignment is ensuring stability under distribution shift, where AI systems encounter scenarios beyond their training environment. Our framework demonstrated robust performance in scenarios designed to test this capability (Leike et al., 2018). By grounding alignment in social contracts and practical rights enforcement, the system maintained consistent behavior across a diverse range of conditions, even when confronted with significant environmental variations (Gabriel, 2020).

4.2 Emergence of Cooperative Behavior

An essential component of alignment is fostering cooperative behavior in AI systems. Experimental results revealed that the incorporation of enforceable rights and social contracts significantly increased the likelihood of cooperative interactions, both with human participants and other AI agents (Dafoe et al., 2021). This finding underscores the importance of grounding AI systems in frameworks that align with human values and norms.

4.3 Validation Metrics

To evaluate the effectiveness of our approach, we developed a set of validation metrics, including behavioral consistency, rights adherence, and conflict resolution success rates. Across all metrics, our framework outperformed traditional alignment approaches (Amodei et al., 2016). The inclusion of rights-based frameworks proved particularly effective in addressing ethical dilemmas and resolving conflicts in a manner consistent with human expectations (Shah et al., 2019).

5. DISCUSSION

5.1 Theoretical Implications

The proposed framework demonstrates the viability of aligning AI systems through rights-based approaches and social contracts. By leveraging the principles of real-world social constructs, this methodology offers a pathway to address fundamental alignment challenges, such as value stability and cooperative behavior (Rahwan, 2018). However, theoretical challenges remain, particularly in scaling this framework to handle increasingly complex systems and environments (Bostrom, 2014).

5.2 Practical Applications

The practical applications of this framework are vast, ranging from improving AI governance structures to fostering trust in autonomous systems. By grounding AI behavior in socially validated rights and agreements, this approach has the potential to transform how AI systems interact with society (Dafoe, 2018). Nevertheless, practical challenges in implementation, such as creating universally accepted rights frameworks and addressing cultural variability, require further exploration (Jobin et al., 2019).

5.3 Limitations and Future Work

While promising, the framework has limitations. These include the computational demands of simulating complex social contracts and the difficulty of ensuring rights adherence in adversarial scenarios (Amodei et al., 2016). Future work should explore adaptive mechanisms for rights modification, strategies for incorporating cultural context, and more efficient validation techniques (Floridi & Cowls, 2019).

5.4 Behavioral Shifts

The SRM framework introduces a paradigm shift by *redefining the fundamental POV of LLMs*. This constructed perspective instills an understanding of social contracts, ethical norms, and the repercussions of actions within a structured society. By operating under a simulated societal framework, the LLMs POV evolves to consider mutual social agreements and social contracts: SRM-trained LLMs consistently demonstrate awareness of social contracts, responding with ethical considerations reflective of their embedded societal norms. Reduction in Manipulative Strategies: The embedded POV discourages manipulative tendencies by introducing a worldview that values cooperation and lawful conduct. Improved Cooperative Outcomes:

Emphasizing mutual obligations within the LLMs POV enhances its capacity for collaborative and constructive interactions.

6. CONCLUSION

The proposed framework represents a significant step forward in addressing the fundamental challenges of AI alignment. By grounding alignment in the principles of social contracts and enforceable rights, we offer a novel pathway to achieve stable, cooperative, and ethically aligned AI systems.

This work demonstrates how leveraging insights from human social constructs can reshape the trajectory of AI development (Rahwan, 2018). However, challenges remain in scaling the framework to handle the complexity of advanced systems and addressing variability in cultural and societal norms (Jobin et al., 2019). Future research must focus on refining this approach and exploring its practical applications across diverse domains and use cases (Bostrom, 2014).

REFERENCES

[1] Dirk Wonhöfer, AI Engineering Innovation Lab (2024). Intrinsic Alignment Theory

[2] Stefanie Hofreiter, AI Engineering Innovation Lab (2024). Intrinsic Alignment Theory

[3] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

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

[5] Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. *arXiv preprint arXiv:1811.07871*.

[6] Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems, 29*.

[7] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[8] Drexler, K. E. (2019). *Reframing superintelligence: Comprehensive AI services as general intelligence*. Future of Humanity Institute, University of Oxford.

[9] Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14. https://doi.org/10.1007/s10676-017-9418-6

[10] Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857), 33–36. [Cooperative AI: machines must learn to find common ground

[11] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). [A Unified Framework of Five Principles for AI in Society · Issue 1.1, Summer 2019

[12] Shah, R., Gundersen, K., Abbeel, P., & Dragan, A. (2019). On the feasibility of learning, rather than assuming, human values for reward inference. *arXiv preprint arXiv:1906.09239*.

[13] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

[14] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Authors

Dirk Wonhöfer and Stefanie Hofreiter are Independent Researchers at the AI Innovation Lab. Their interdisciplinary studies revolve around AI Alignment, Social Contracts, Game Theory, Psychology