


RESEARCH ARTICLE

Reversing the logic of generative AI alignment: a pragmatic approach for public interest

Gleb Papyshev 

Division of Social Science, The Hong Kong University of Science and Technology, Kowloon, Hong Kong, China
Email: gleb@ust.hk

Received: 14 July 2024; **Revised:** 19 December 2024; **Accepted:** 06 February 2025

Keywords: AI alignment; constitutional AI; pragmatism; public interest; reinforcement learning with human feedback

Abstract

The alignment of artificial intelligence (AI) systems with societal values and the public interest is a critical challenge in the field of AI ethics and governance. Traditional approaches, such as Reinforcement Learning with Human Feedback (RLHF) and Constitutional AI, often rely on pre-defined high-level ethical principles. This article critiques these conventional alignment frameworks through the philosophical perspectives of pragmatism and public interest theory, arguing against their rigidity and disconnect with practical impacts. It proposes an alternative alignment strategy that reverses the traditional logic, focusing on empirical evidence and the real-world effects of AI systems. By emphasizing practical outcomes and continuous adaptation, this pragmatic approach aims to ensure that AI technologies are developed according to the principles that are derived from the observable impacts produced by technology applications.

Policy Significance Statement

This article advocates for a pragmatic, public interest-driven approach to AI alignment, challenging the theoretical underpinnings of current methods like RLHF and Constitutional AI. The proposed strategy emphasizes empirical evidence and practical outcomes, prioritizing real-world impacts over abstract ethical principles. By analyzing instances where AI systems have failed or caused harm, developers can establish actionable, experience-based principles in a bottom-up manner. This approach mitigates reward hacking risks and aligns AI development with societal needs. This democratic process ensures accountability and reflects evolving public values, providing a robust framework for creating ethically responsible and effective AI systems that serve the public good.

1. Introduction

The rapid advancement of artificial intelligence (AI) technologies has brought about significant benefits and transformative potential across various sectors. However, alongside these advancements come pressing concerns regarding the alignment of AI systems with societal values and the public interest. Traditional approaches to AI alignment, such as Reinforcement Learning with Human Feedback (RLHF) and Constitutional AI, often emphasize adherence to pre-defined ethical principles and high-level norms. Despite their popularity, these methods face challenges in ensuring that AI systems consistently act in ways that benefit the broader public.

The main focus of this article is to discuss the theoretical underpinning behind the popular approaches to AI Alignment from the perspective of the pragmatist philosophy and to describe an alternative conceptual model that is inspired by the ideas of the public interest theory and legal pragmatism, with the key proposition being the introduction of inductive evidence-based principle derivation instead of the deductive normative rule-applications as a guiding principle.

Constitutional AI is an approach to value alignment that aims to ensure AI systems behave in accordance with a predefined set of principles or “constitution.” Instead of relying solely on human feedback to shape the AI’s behavior, constitutional AI embeds these principles into the core of the AI system’s decision-making process (Bai et al., 2022). Anthropic is the company championing this approach to alignment.

The key idea behind constitutional AI is to create a framework of rules, values, and constraints that guide the AI’s actions and ensure it remains aligned with human values. These principles are chosen by the developers of the models and may include concepts like respect for human rights, transparency, fairness, accountability, and the promotion of beneficial outcomes for humanity (Conitzer et al., 2024).

Once the constitutional principles are established, they are hardcoded into the AI system’s architecture. This means that the AI is inherently bound by these principles and cannot violate them, even if doing so would lead to more optimal outcomes according to its objective function. The AI’s decision-making process is constrained by the constitutional framework, ensuring that it operates within the boundaries of what is considered acceptable and aligned with human values (Kundu et al., 2023). While Constitutional AI’s reliance on hardcoded principles ensures a level of safety and predictability, it also introduces rigidity that may prevent AI systems from dynamic adaptation, which plays a crucial role in mitigating potential pitfalls of unchecked goal maximization, such as extreme risk-taking behaviors.

RLHF, on the other hand, is a different technique used to fine-tune language models like GPT-4. This approach is especially prominent in OpenAI’s work (Ouyang et al., 2022). It involves collecting a dataset of human-written demonstrations of desired output behavior on prompts, training a supervised learning baseline on this data, and then collecting a dataset of human-labeled comparisons between model outputs on a larger set of prompts. A reward model is then trained on this dataset to predict which model output the human labelers would prefer. Finally, the supervised learning baseline is fine-tuned to maximize the reward predicted by the reward model, aligning the language model’s behavior with the preferences of the human labelers and researchers (Ouyang et al., 2022). By incorporating human preferences into the training process, RLHF aims to align the language model’s behavior with the values expressed by the human labelers.

Situated at the intersection of AI and human–computer interaction, RLHF introduces a human-in-the-loop approach to the classical reinforcement learning paradigm, creating an opportunity for the potential alignment of AI agents’ goals with the ethical values expressed by humans (Kaufmann et al., 2023). However, to achieve a more representative alignment process, the inclusion of various stakeholders is necessary. This inclusivity can lead to the expression of conflicting values, potentially compromising the model’s performance (Conitzer et al., 2024). Furthermore, alignment is often rather cosmetic and superficial, bringing only stylistic changes to the model’s performance (Lee et al., 2024). In safety-critical domains, the problem of low-confidence feedback from people may accelerate potential incidents caused by the application of large language models (Chaudhari et al., 2024). Newer techniques of generating feedback by AI systems themselves and then incorporating it into the system’s design, while having shorter annotation cycles and lower costs (Li et al., 2024), can suffer from the subtle incorporation of non-aligned information in the system’s design.

This article critiques the conventional alignment frameworks through the philosophical lenses of pragmatism and public interest theory. By examining the limitations of RLHF and Constitutional AI, it proposes an alternative approach grounded in pragmatism. Pragmatism, with its focus on practical outcomes and real-world effects, offers a robust foundation for rethinking AI alignment.

The subsequent sections of this article will explore the concept of pragmatism and its application in governance, delve into public interest theory, and analyze RLHF and Constitutional AI through the perspectives of legal theory. The discussion will culminate in proposing a reversed logic for AI alignment,

one that prioritizes empirical evidence and the practical impacts of AI systems. By adopting this pragmatic approach, the goal is to develop AI technologies that are not only ethically responsible but also aligned dynamically with the evolving impacts on society, thereby ensuring they serve the collective well-being and social good.

2. Defining pragmatism

Pragmatism is a philosophical school developed by American philosophers Charles Peirce, William James, and John Dewey between the late 19th century and the Second World War (Ansell and Boin, 2019). Later contributions to the development of this philosophical approach were made by adding certain features of other philosophical frameworks to pragmatism, such as Richard Rorty's postmodern pragmatism, (Allen, 2008), Hilary Putnam's pragmatic realism (Sosa, 1993), and Sydney Hook's alignment of pragmatism with Marxism (Sidorsky and Talisse, 2018). Though debatable, the wide-ranging philosophical ideas of Jurgen Habermas can also be considered along the line of pragmatism (Sager, 2007).

Among the classical pragmatists, Charles Peirce was the first one to define pragmatism as a new philosophical movement (Legg and Hookway, 2021). Peirce establishes a pragmatist maxim in the article "How to Make Our Ideas Clear," which says: "Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of those effects is the whole of our conception of the object" (Buchler, 1955). Classical pragmatists regarded this maxim as the basic principle of their philosophy and developed other ideas (such as the pragmatist theory of truth) to apply this maxim (Putnam, 1995; Pihlström, 2008).

Even though Peirce was loyal to this definition of the pragmatist maxim, he elaborated on it in his other writings by reinterpreting it as: "Pragmatism is the principle that every theoretical judgment expressible in a sentence in the indicative mood is a confused form of thought whose meaning, if it has any, lies in its tendency to enforce a corresponding practical maxim expressible as a conditional sentence having its apodosis in the imperative mood" in 1903, and as "The entire intellectual purport of any symbol consists in the total of all general modes of rational conduct which, conditionally upon all the possible different circumstances and desires, would ensue upon the acceptance on the symbol" in 1905 (Hookway, 2008, 2012).

In a philosophical sense, this maxim should not be treated as a theory of meaning but rather as a tool for improving inquiry, and particularly for determining the significance of one intellectual concept over another (Olszewsky, 1983), based on the investigation of the practical consequences of this concept (Cherryholmes, 1994). Thus, the core idea of the maxim is that by identifying the effects that an intellectual concept is having, it is possible to make the content of this concept fully explicit (Hookway, 2009), or to put it differently, "our conception of an object or situation is the sum-total of the practical bearings we conceive it to have" (Ulrich, 2007).

Other pragmatist philosophers further elaborated on this maxim, among whom William James was one of the first and most prominent contributors to the development of pragmatism. Williams provides his interpretation of the pragmatist maxim by saying: "To attain perfect clearness in our thoughts of an object, then, we need only consider what conceivable effects of a practical kind the object may involve—what sensations we are to expect from it, and what reactions we must prepare. Our conception of these effects, whether immediate or remote, is then for us the whole of our conception of the object, so far as that conception has positive significance at all" (James, 1975).

To illustrate this understanding of the maxim, James provides an example of a philosophical dispute revolving around a man trying to see a squirrel on the opposite side of the tree. But, no matter how fast this man goes, the squirrel moves faster, maintaining a distance between it and a man so that it is never seen. James then asks, "Does the man go round the squirrel or not?". He provides a pragmatic explanation of this philosophical dispute. "James proposed that which answer is correct depends on what you 'practically mean' by 'going round.' If you mean passing from north of the squirrel, east, south, then west, then the answer to the question is "yes." If, on the other hand, you mean in front of him, to his right, behind him, to

his left, and then in front of him again, then the answer is ‘no.’ After pragmatic clarification disambiguates the question, all dispute comes to an end” (Legg and Hookway, 2021).

The idea of truth becomes central to the writings of James, as his argumentation was predominately targeted against the philosophy of idealism. In contrast to idealists, in his understanding, pragmatist truth had to be understood as coherence between ideas (thoughts) and reality (Thayer, 1977). As James puts it, “Any idea upon which we can ride, so to speak; any idea that will carry us prosperously from any one part of our experience to any other part, linking things satisfactorily, working securely, simplifying, saving labor; is true for just so much, true in so far forth, true instrumentally. This is the ‘instrumental’ view of truth taught so successfully at Chicago, the view that truth in our ideas means their power to ‘work,’ promulgated so brilliantly at Oxford” (James, 1975). This instrumental view of truth allows him to mark some ideas as ‘true’ or ‘good’ if they have significance for specific actions (Brandom, 2008), essentially labeling truth as something that ‘works’ (Cormier, 2000). Thus, the idea becomes ‘true’ if it is made true by events and verified by its application to the reality (Capps, 2019).

This view on what constitutes truth has established the image of pragmatist truth as something that can justify utilitarian moral theory (Marchetti, 2010) or post-truth (Forstenzer, 2018; de Freitas Araujo, 2020), which is reflected in some pragmatist philosophers not truly embracing democratic values, but it does not mean that it can justify any actions without taking into account how much moral effort they require and whether their effect is positive or negative, good or lousy (Putnam, 2006). Along these lines, Bertrand Russell famously commented that William James is committed to the truth that Santa Claus exists, but “at best, James is committed to the claim that the happiness that belief in Santa Claus provides is truth-relevant. James could say that the belief was “good for so much,” but it would only be “wholly true” if it did not ‘clash with other vital benefits’ (Legg and Hookway, 2021).

The instrumental approach to pragmatist philosophy was further actively developed by John Dewey (Welchman, 1989), who saw philosophy primarily as a method for inquiry, an instrument for the clarification of the course of life (Upin, 1993). In its broad sense, instrumentalism refers to the idea that knowledge originates in human practice and that its veracity can be determined by examining the application of concepts to concrete problems (Shuklian, 1995). Thus, Dewey believes that the primary function of thought is to make indeterminate situations more determined, which requires planning and inquiry (Wible, 1984) so that unsettled situations can be resolved by utilizing knowledge for human needs (Rudolph, 2005).

The practice of inquiry is of particular importance in the philosophy of Dewey as he defines it as “the controlled or directed transformation of an indeterminate situation into one that is so determinate in its constituent distinctions and relations as to convert the elements of the original situation into a unified whole” (Dewey, 2018). According to Dewey, inquiry begins with a feeling of unique doubtfulness, followed by a formulation of the problem, a hypothesis about its resolution, reasoning through the alternatives, and finally, making a decision about an act (Hildebrand, 2021). Though not without critique (Dicker, 1971), this approach to rational inquiry views it as relating means with consequences, as part of the practical means-end reasoning, and not one of the abstract principles (Garrison, 1999).

Truth, as understood by Dewey, is then not a fixed absolute that has to be discovered but rather a process of inquiry itself; meaning that to claim that something is “true” is to acknowledge that it is a reliable resource built on which a further inquiry can be undertaken (Hildebrand, 2021). Thus, in the pragmatist tradition, while Peirce and James had developed the pragmatist theory of what truth is, Dewey instead focused his attention on highlighting what truth does by showing that it is a constant process of refining our understanding of certain phenomena through further investigation of its practical effects (Capps, 2018).

Richard Rorty further elaborated on the idea of pragmatist truth. He believed that concepts such as truth, objectivity, or reality lack practical purposefulness because they cannot legitimate the standards of the warrant. That truth should not be the goal towards which people should strive (Ramberg and Dieleman, 2021). This rather postmodernist take on pragmatism (Allen, 2008) is related to the instrumentalist view of truth developed by Dewey because, according to this view, the truth can only be recognized as something that is accepted and endorsed as a standard for a certain period in a particular

community (Legg and Hookway, 2021). Thus, truth should not be understood as something that can be theorized about but should rather be seen as something that brings “good” through action (Williams, 2003). What distinguishes this view of truth from postmodern relativists is the need for the idea to have practical bearings on the world to be justified as “true” or “good” (Kumar, 2005).

The idea of truth as an absolute that should be found in the world through the process of inquiry is then seen by Rorty as impractical because “pragmatists think that if something makes no difference to practice, it should make no difference to philosophy” (Rorty, 1995); just as what the concept of truth is in his understanding. This opinion is, perhaps, best described by Richard Rorty himself in one of the interviews, where he says that “truth with a capital T is sort of like God. There is not much you can say about God... Contemporary pragmatists tend to say the word “true” is undefinable, but none the worse for that we know how to use it, we don’t have to define it” (White, 2008).

The idea that an intellectual object is the sum-total of its practical bearings on the world, the perception of the process of inquiry as the transformation of something indeterminate into something determined, and the view that truth is something that has positive practical effects on the world and is accepted as a standard (though which can be revised over time) are among the key ideas that differentiate this philosophical school from others. The following section will discuss how these pragmatist ideas were further developed and applied to governance studies.

3. Pragmatism and governance

The theories introduced by the classical pragmatists were later adopted by Harold D. Laswell for the development of the discipline of policy sciences (Kaufman-Osborn, 1985)—an approach to investigate the process of policymaking so that the knowledge created through it can be applied to improve this process (Dunn, 2018b). This pragmatic take on the study of policy process has become widely adopted in different policy sub-domains, including foreign policy (Tjalve, 2013), urban policy (Bridge, 2008), public health (Iyasu, 2009), education (Stone, 1998), and environmental policy (Minteer et al., 2004).

By transforming the ideas of Peirce’s pragmatic maxim and Dewey’s instrumentalism, Laswell has developed his maxim for the discipline of policy sciences, which “should be conceived as knowledge of the policy process and the relevance of knowledge in the process” (Laswell, 1970). This pragmatic policy mantra can be understood as an approach that looks at how policy decisions affect and are affected by the social contexts in which they take place, which include participants, their perspectives and beliefs about the issue, as well as their strategies for action, and the outcomes and effects that they bring (Dunn, 2019).

Dewey’s instrumentalism inspired Laswell to develop his functionalist decision theory (Dunn, 2018a), which perceives the policy process as a set of purposive acts toward values of human dignity, enlightenment, wealth, power, and rectitude. It includes intelligence (gathering and disseminating information), promotion (propaganda), prescription (institutionalization of norms), invocation (enforcement of the prescriptions), application (written documentation), termination (cancellation), appraisal (legal responsibility for the policy assessment) (Laswell, 1956). This cyclical understanding of the government as a process, rather than a set of institutions interacting with each other, has become popular in the disciplines of public policy and public administration, where the idea of policy cycles is one of the main theoretical frames through which the study of the policy process is conducted (Bridgman and Davis, 2003; Fischer and Miller, 2006).

The idea of the pragmatic maxim has been especially relevant for the study of public administration, where the bureaucrats are dealing with constantly changing real-world situations and have to adapt their behaviors on the go to reach their objective, which is distinctively different from a more distanced and abstract study of public policy (Shields, 1996). Following the ideas of Dewey, problematic situations that bureaucrats encounter in their activities require them to conduct an inquiry to find a satisfying solution to their problem (Shields, 2003)—which is not necessarily the “true” solution to the problem, but rather something that is “useful” and “works” at a particular point in time (Miller, 2004). Thus, the pragmatic evaluation of the results of specific actions is not conducted from an objective vantage point but instead

considers them through an ethical stance of classical pragmatism—what practical results in the everyday life of individuals are brought by it (Shields, 2008).

While the pragmatic view on the discipline of public administration can create specific theoretical explanations for the events that are taking place in the field, pragmatism also offers a strategy for the actual bureaucrats who tend to work with fixed concepts (policies, principles, laws) often expressed in vague and general terms, and apply them in a fluid and ever-changing circumstances and unusual situations (Hildebrand, 2005). Thus, the theorization of these processes should also take place through the understanding of the actional situations in which this process takes place because the uncertainty of human behavior and practical actions, even in organized and routinized activities, is never an easy task to consider and comprehend, especially in the production of theories that include this element of uncertainty (Chailleux and Zittoun, 2022).

As such, the ideas of experimentalist governance and the conduct of policy experiments are one of the ways through which the policymaking process can become more pragmatic, as the conduct of policy experiments resembles the idea of inquiry in the classical pragmatist sense (Overdevest et al., 2010). Experimentalist governance can reduce uncertainty by using small-scale trials and errors that would eventually shed light on the policy design that “works,” resulting in risk reductions and policy innovation (Nair and Howlett, 2016).

Another application of the pragmatist approach to policy experiments is regulatory sandboxes—“safe” controlled legal environments monitored by the regulatory authority, where innovative companies can be exempted from certain regulatory obligations for a determined period while granted the opportunity to test their solutions on real customers in actual markets (Ringe and Ruof, 2018). Though regulatory sandboxes first appeared in the FinTech industry (Allen, 2019), currently this approach has been adopted by many countries primarily for regulating emerging technologies, such as AI (Truby et al., 2021), and have been considered to be a successful way of managing regulation around innovative technologies (Goo and Heo, 2020).

The pragmatic look at the policy process and its administration must concentrate primarily on the gap between policy goals and their actual impact. As such, the perception of the policy can be constructed through understanding the difference between its intended and actual effects (Rigby, 2005)—though, in some more radical interpretations, it can also be seen as a framework that prioritizes materialistic concerns over moral principles under reality constraints (Lee, 2010) and holds no moral or political positions (Knight and Johnson, 1999).

Though pragmatism indeed may seem to be a valueless approach to addressing a policy issue, it can also be just another way of looking at how moral values are incorporated into policy actions—essentially reversing the logic of principle-based ethics to a more practical method of deriving ethical values from practice (Minteer et al., 2004). Under this understanding, principles are not seen as existing abstract entities brought to life via specific actions. Instead, the practice of doing something can constitute value-making (Oldenhof et al., 2022). Thus, the lived practices are becoming the policy tools that are used for the implementation of policy ideas, where the values and principles associated with these actions can provide feedback to the overall redesign of the policy itself; “the alternative—of prematurely foreclosing necessary social debate on the question of how we might learn to live together—is perilous indeed to long-term social development” (West and Davis, 2010).

A pragmatic look at the ethical issues related to policies reverses the idea that policies should be based on certain moral principles and act as mediums through which these principles are introduced to society; instead, it argues that policies should be looked at through the prism of the practical value that they bring to the society (Pouryousefi and Freeman, 2021). Thus, the study of policies in action should be based on understanding actionable knowledge, the interconnectedness between experience, knowing, and acting, and base inquiry on experiential interpretations (Kelly and Cordeiro, 2020).

4. Public Interest Theory

The Public Interest Theory represents an approach to the regulatory frameworks that shape government intervention in various sectors. It is predicated on the notion that such interventions are not mere exercises of authority but are instead crafted with the overarching goal of protecting and promoting the collective well-being of society. This theory contends that the state's regulatory activities are a response to certain market failures, such as monopolies, externalities, or information asymmetries, which, if left unchecked, could lead to suboptimal outcomes for the people (Bozeman, 2007).

At the heart of the Public Interest Theory lies the assertion that government regulations should play a pivotal role in correcting market inefficiencies that thwart fair competition and impede the equitable distribution of resources (Alexander, 2002). The theory underscores the need for regulations to create conditions where markets operate with maximal efficiency, characterized by open and voluntary exchanges that are free from distortions that could disadvantage consumers or other market participants.

This theoretical framework elevates the concept of public welfare above individual or private interests, suggesting a governance model where individual pursuits must be harmonized with the collective good. In this sense, the Public Interest Theory does not dismiss the importance of individual interests but rather integrates them within a broader societal context, ensuring that the pursuit of private goals does not come at the expense of public welfare. It posits a normative vision that encourages regulators to make decisions that are not solely driven by economic considerations but also by the potential to improve societal welfare.

Moreover, the theory suggests that when policymakers align their regulatory actions with the public interest, they enhance the legitimacy and effectiveness of governance. But this is based on the idea that what constitutes the "public interest" or the "public good" is always clear for the designers of the intervention, which is not always the case (Rex, 2020).

The concept of "public good" is a term with deep roots in social and political theory, often invoked in discussions about initiatives that aim to serve the collective interests of society. However, its definition is fraught with complexity and ambiguity. This is particularly true when it comes to emerging technologies like AI, where the term "public good" can be interpreted in multiple ways (Züger and Asghari, 2023).

Pragmatism, as a philosophical doctrine, suggests that the truth of an idea is determined by its practical effects and outcomes. This perspective is echoed in the Public Interest Theory, where policies are evaluated not just on their intentions but on their real-world effects on the public. The pragmatic approach in policy-making underscores the importance of empirical evidence and the actual impact of policies when determining their value, rather than relying solely on theoretical or ideological considerations.

From the pragmatic viewpoint, public interest is not defined a priori by some moral or ethical standards; rather, it is determined through the observation of the outcomes that policies have on society. If a policy leads to the betterment of the public's welfare, it can be said to be in the public interest. In this way, pragmatism informs the Public Interest Theory by emphasizing the need for policies to be judged by their practical and tangible benefits to society.

Similar division along the lines of principles and practices can be found in the field of legal scholarship, where there are two ways of approaching the subject: a formalist way of assigning fixed meaning to legal principles and their correspondence to things and a realist approach which is more concerned with placing the legal doctrine within a broader socio-political landscape of our societies and deriving the meaning of the legal intervention through the examination of empirics (Unger, 2021). The next section analyzes approaches of RLHF and Constitutional AI through these lenses adopted from legal scholarship and provides a critique based on the philosophy of pragmatism and the public interest theory.

5. AI alignment through the Lens of Legal Theory

Legal formalism perceives law as an "immanent moral rationality" (Unger, 1983) and argues that the law is autonomous from other social forces and establishes a legal theory implicit within the legal doctrine itself, which elaborates on itself from within (Weinrib, 1993). The key idea of legal formalism is that law is an internally coherent phenomenon, which is different from politics that it is surrounded by; because law

inhibits a form of rationality other than the political or ideological contests, it has distinctive immanence to the legal material which it interrogates, and develops a normative theory solidifying the tradition to form an intelligible moral order (Weinrib, 1993). Legal formalism approaches legal practice as a deductive exercise to derive and judge the outcome of a case based on authoritative principles, essentially allowing the commentator to view the outcome of the case as either “correct” or “incorrect” in a nearly mathematical sense (Posner, 1986).

By seeing formalism as a scientific theory of law, the proponents of this theoretical stance argue that all judges have the same psychology when they are making a decision—as there is always the right decision about a particular case—and, thus, the doctrine is making the verdict using the judges as its avatars; or that the judges follow specific black-letter rules when making their decision, which is often based on the decisions made by other judges in similar situations (Troop, 2018). As such, the formalist approach views legal texts as tools for solving social problems, the purpose of which is implicit in their form of self-sufficient objects or as a window to the ephemeral manifestation of a more absolutist conceptual order (Malkan, 1997).

For classical formalists, the issue can never be resolved by looking at practical efficiencies. Instead, the correct answers must be deduced from the fundamental principles - always viewing legal thinking as a deductive rule-applying (Pildes, 1999). Thus, argues that to the extent possible, the application of the legal rules in society should be conducted mechanically unless there is an extreme deviation from the norm (Peters, 2014). Being a rules-based decision-making paradigm, formalists treat the doctrine and tradition as absolute values allowing little flexibility and innovation, which requires formalists to stick to the pre-existing principles and violate informal moral norms present within the society and derived from standard practices, which are not directly aligned with the absolutist principles (Farber, 2015).

Both RLHF and Constitutional AI techniques follow a similar logic to legal formalism, where they are built upon an assumption that high-level abstract ethical norms exist in reality and can be brought to life through deductive rule-applying. While Constitutional AI strictly follows certain principles that are chosen by its developers, RLHF also suggests the desirable values that should be accentuated in the model’s behavior with the help of human evaluators of its outputs.

This creates a power disparity between the model developers, who choose the abstract values with which to align their systems, and users who are subjected to such systems, often without transparency about the chosen principles. This situation goes against the logic of pragmatist ideas and the public interest theory, as it creates a rigid system based on a fixed number of high-level ethical principles chosen by a small group of people, instead of creating an inclusive and agile system.

As such, both RLHF and Constitutional AI exhibit significant shortcomings, particularly when evaluated against the philosophy of pragmatism, which emphasizes practical consequences and real-world applications. While RLHF and Constitutional AI aim to align AI systems with human values, they fall short of serving the broader public interest due to their lack of transparency and inclusivity. While supporters of this approach would argue that secrecy is needed to prevent gaming the system, it is still based on trusting the system designers with ethical choices, instead of relying on a more democratic procedure.

One major issue with RLHF is its opacity. The specifics of human feedback and the reinforcement signals used in training are not fully disclosed to the public. This secrecy not only prevents public scrutiny but also stifles broader understanding and engagement. The concentration of decision-making power in the hands of a few individuals at OpenAI further exacerbates this problem. Such centralization is antithetical to the pragmatic approach, which advocates for practical, inclusive solutions that emerge from widespread participation and transparent processes.

Similarly, Anthropic’s Constitutional AI approach, which involves hardcoding specific values and principles into AI systems, lacks the flexibility and adaptability that pragmatism demands. The principles guiding these AI systems are determined by a small group of researchers with limited public input or oversight. This process is opaque, making it difficult for the public to understand how these values are selected and embedded. Pragmatism calls for solutions that can evolve with changing societal norms and values, but hardcoded principles may become outdated and misaligned with the dynamic needs of society.

Both RLHF and Constitutional AI risk perpetuating biases due to their limited and non-representative input sources. The human feedback in RLHF often comes from a narrow sample of individuals, potentially leading to an AI that reflects and amplifies specific perspectives while marginalizing others. Constitutional AI faces a similar risk, as the values embedded in the AI may reflect the biases of the researchers involved. This selective input fails to align with the pragmatic emphasis on broad, inclusive, and practical engagement with diverse societal perspectives. For example, the list of principle categories used by Anthropic includes: Principles Based on the Universal Declaration of Human Rights; Principles Inspired by Apple’s Terms of Services; Principles Encouraging Consideration of Non-Western Perspectives; Principles Inspired by DeepMind’s Sparrow Rules; Principles from Anthropic Research (Bai et al., 2022). Yet the reasons why the principles were derived from certain frameworks and not others remain unclear. Though there are attempts to democratize constitutional AI through the engagement of the wider public for constitution drafting, so far, such involvement has been limited to 1000 Americans only, leaving the vast majority of the model’s potential users outside of value alignment conversation (Anthropic, 2023). This limited engagement leaves out the vast majority of potential users, failing to capture the full spectrum of societal values and needs. Pragmatism would advocate for a more extensive and iterative process of public involvement to ensure that AI systems are truly aligned with the evolving values of society.

The concentration of power in the development of these AI systems is another significant issue. When a small group of individuals or organizations, often from similar backgrounds, determine the guiding values and principles, there is a risk of reinforcing existing power structures and inequalities. This market concentration is a key feature of the current AI landscape globally, where only a select number of organizations are capable of developing the most advanced systems. This scenario runs counter to the pragmatic approach, which values solutions that arise from broad-based, inclusive participation and are responsive to the practical needs of all societal segments.

6. Reversing the logic of AI alignment

The problems highlighted above make both RLHF and constitutional AI approaches potentially not acting in the public interest when used to align AI systems with certain values and principles. The critique of the existing alignment techniques is not new, as several studies have addressed this issue from different angles. As such, social choice theory becomes both a critique of the current approaches that do not account for the democratic inputs for value alignment due to the lack of the institutional architecture (and potentially the desire of the system developers) to allow for it (Prasad, 2018), as well as a supporter of the principles-based value alignment if a system that accounts for conflicting opinions in a democratic manner exists in an ideal world (Conitzer et al., 2024). Nevertheless, even if such an institutional environment was designed, either with the involvement of regulators or through other means, the social choice approach still suffers from the potential to be manipulated or “gamed” through the inherent loopholes in its design, such as optimizing the values for the average preference of the group, but limiting whose opinions are included in this group (Baum, 2024). Furthermore, the rigidity of the social choice theory does not allow for a more dynamic and radical value alignment through personalization as suggested by the proponents of epistemic agency (Huang et al., 2024).

Nevertheless, the pragmatic value alignment proposal does not disregard the ideas of social choice theory and epistemic agency completely. While the democratization of the pool of participants whose opinions regarding value alignment should be increased—something that the pragmatists would agree with—the alignment according to predefined ethical principles without any grounding in reality is something that contradicts the pragmatist maxim. As such, the pragmatist value alignment proposal accentuates the need to systematically study the negative effects (incidents) that AI systems are already having on the real world (e.g. the OECD has collected over 15000 incidents in its database (*The OECD Artificial Intelligence Policy Observatory*, n.d.)) and to derive the higher-level categories of risks related to ethical principles. This will allow aligning the target behavior with the real threats that are already caused by AI applications and potentially identify the categories that need to be targeted during the

alignment process, reversing the logic from ad-hoc normativity to post-hoc empiricism, allowing expansion of the instruments used for alignment from just formalist principle application, which fails short to account for practical effects.

The philosophy of pragmatism, on the other hand, combined with public interest theory, offers a foundation for designing an alternative approach to the AI alignment problem. This alternative focuses on the real-world effects of AI systems, ensuring these effects align with the public interest, by building upon the critique that Legal Realism provides for Legal Formalism. In other words, reversing the logic of AI alignment from deductive rule-applying to inductive rule-deriving.

The legal realists criticize the formalist stance that law is objective, extrinsic to the social forces, and superior to politics. Instead, they emphasize the fact that the act of judging is very heavily influenced by the values that the judges have (also applies to evaluators in RLHF and developers in Constitutional AI)—thus, understanding the working of the law within the society with all its limitations and flaws, as well as acknowledging the fact that judges can influence the execution of the law and can be influenced by their own moral and political views and biases (Tamanaha, 2008).

On the surface level, the realists think that any decision made by a judge can be attributed to the judge's political, social, or economic values—or, to put it differently “what the judge had for breakfast”; a more sophisticated look at this problem, however, suggests that the indeterminate nature of law itself leaves the judges with too many principles, which “are too numerous, too imprecise, and too malleable” to choose from, allowing them to freely decide on the verdict in the absence of adequate guidance (Newman, 1984). The realists argued, first, that legal rules often contain abstract and complex defined concepts, such as “reasonableness” or “duress,” letting judges interpret these concepts in their way and not be able to apply them mechanically; second, that it is tough to derive a principle from a case because each case can be read in two ways: broadly to establish a general rule applicable to many different situations, or narrowly focusing on the particular facts of the case; third, that the indeterminacy of abstract concepts and manipulability of the case interpretations allows for the appeal to contradictory rules to decide a contested case (Singer, 1988).

Instead of treating law as a set of given axioms, the realists see it as both the instrument of the government to establish and execute its power within the society but also as a means through which it is possible to deduce the principles from what is good for the community based on the realist examination of the empirics (Posner, 1986). The most radical form of legal realism, the so-called “rules skeptics” argue that real law only consists of the past judicial decisions, claiming that there is never anything determinate in the law apart from the past decisions and that its predictive power is nearly non-existent—essentially proposing a legal reinterpretation of the pragmatist maxim: “a judicial decision is justified when, but only when, it serves (perhaps to the maximum degree possible) the interests of those who will be affected by the decision” (Lyons, 1981).

The realists argue that law should be studied as “it is” because an empirical investigation of the practice of law can reveal that legal rules can differ from what the courts do in practice. The presupposed effects of the legal doctrine on society should be studied not normatively but through a careful examination of their practical impact on society (Williamson, 1996). This allows the realists to acknowledge that not every aspect of social life is governed through top-down legal principles, but rather that many domains are governed through bottom-up non-legal systems of norms, which often can contradict each other because the actual situation on the ground can be in stark opposition to the legal principles, yet still be accepted by the society as the right way of doing things (Radin and Wagner, 1999). Thus, the distinction between the “paper rules” and the “real rules” has become one of the central arguments of realists (Schauer, 2012).

This brings us to the key idea behind the pragmatic approach to AI alignment which is to concentrate on the actual outcomes produced by AI systems and ensure these outcomes do not contravene the public interest. This task is challenging if the systems are constructed with the logic of incorporating high-level ethical principles or teaching the systems an “ideal” behavior. When built with this logic, AI can optimize its reward functions to “hack the system” by exploiting loopholes in the objective's definition (Krakovna et al., 2020).

To align systems pragmatically and maximize public interest, a reversed approach to alignment logic should be introduced. This approach would focus on instances when AI systems have caused harm or exhibited unwanted behaviors in the real world and derive pragmatic principles from these incidents. By examining real-world cases where AI systems acted in ways that were not aligned with the public interest, system designers can extract pragmatic principles from the violated norms. This method ensures that AI systems are aligned with principles derived from actual negative incidents, thus grounding the alignment process in practical, real-world outcomes.

For example, consider an AI-powered content moderation system designed to reduce harmful speech. If aligned using conventional methods, it may apply predefined ethical principles or steer its behavior towards certain population groups' values, but in practice, this could lead to over-censorship, disproportionately silencing marginalized voices. A pragmatic alignment approach would instead analyze real-world cases where the system either failed to block harmful content or unfairly flagged benign speech. By systematically identifying recurring issues and adjusting AI behavior based on empirically observed harms, this method ensures alignment is grounded in actual societal impacts rather than abstract ethical norms.

This new alignment method emphasizes continuous monitoring and adaptive learning, rooted in the philosophical tenets of pragmatism. Pragmatism, which values action and results over abstract theorizing, suggests that knowledge and meaning are contingent upon their practical applications and consequences. By shifting the focus from pre-defined ethical abstractions to the concrete impacts of AI behaviors, this approach embodies the pragmatic insistence on the importance of experience and empirical evidence. The iterative process of assessing and adjusting AI behaviors in response to real-world outcomes reflects the pragmatic belief in the fluidity and adaptability of knowledge, highlighting the necessity for AI systems to evolve in tandem with societal changes.

Furthermore, this alignment strategy aligns with the philosophical underpinnings of public interest theory, which emphasizes the role of collective well-being and social good. By systematically documenting and analyzing instances of AI misalignment, this approach fosters a more transparent and inclusive process, encouraging public participation in the ethical governance of AI. This engagement not only democratizes the alignment process but also reinforces the accountability of AI developers to the broader public. In doing so, it mitigates the risk of privileging the values of a select few over the diverse interests of society. Through this theoretical and philosophical lens, the proposed alignment method offers a robust framework for developing AI systems that are both ethically responsible and practically effective, ensuring their alignment with the dynamic and evolving values of the public interest.

7. Conclusion

This article presents an alternative method for AI alignment grounded in the philosophy of pragmatism and the public interest theory. The key argument presented here states that the two most prominent approaches to AI alignment—RLHF and Constitutional AI—may not be well suited for benefiting the public interest.

Instead, a reversed approach to AI alignment based on pragmatist principles could overcome the shortcomings of the existing approaches by refocusing on instances when AI systems already act against the public interest in real-world settings. This method emphasizes empirical evidence and practical outcomes over abstract ethical guidelines, ensuring that AI development is continually informed by actual experiences and societal impacts. By systematically analyzing incidents where AI systems have caused harm or deviated from intended behaviors, developers can derive actionable principles that directly address these failures. This approach not only mitigates the risks associated with reward hacking but also ensures that the alignment process is grounded in the practical realities of AI deployment.

Furthermore, this pragmatic alignment strategy promotes a more inclusive and transparent process by actively involving diverse stakeholder perspectives. Engaging the public and various interest groups in the ethical governance of AI fosters a more democratic approach to alignment, ensuring that the values and needs of a broad spectrum of society are considered. This participatory model not only enhances the

accountability of AI developers but also aligns the development process with the evolving values of the public interest. By adopting these principles, the proposed alignment method offers a robust framework for creating AI systems that are both ethically responsible and practically effective, ultimately ensuring that AI technologies serve the public good.

Data availability statement. Data availability is not applicable to this article as no new data were created or analysed in this study.

Author contribution. This is a single-author manuscript.

Funding statement. No funding was received to conduct this study.

Competing interest. The author states that there is no conflict of interest.

References

- Alexander ER (2002) The public interest in planning: From legitimation to substantive plan evaluation. *Planning Theory* 1(3), 226–249. <https://doi.org/10.1177/147309520200100303>.
- Allen B (2008) Postmodern pragmatism: Richard Rorty's transformation of American philosophy. *Philosophical Topics* 36(1), 1–15.
- Allen HJ (2019) Regulatory sandboxes. *George Washington Law Review* 87, 579.
- Ansell C and Boin A (2019) Taming deep uncertainty: The potential of pragmatist principles for understanding and improving strategic crisis management. *Administration & Society* 51(7), 1079–1112. <https://doi.org/10.1177/0095399717747655>.
- Anthropic (2023). Collective Constitutional AI: Aligning a Language Model with Public Input. <https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input>.
- Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli D, Li D, Tran-Johnson E, Perez E, ... Kaplan J (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv. arXiv:2212.08073.
- Baum SD (2024) Manipulating aggregate societal values to bias AI social choice ethics. *AI and Ethics*. <https://doi.org/10.1007/s43681-024-00495-6>.
- Bozeman B (2007) *Public Values and Public Interest: Counterbalancing Economic Individualism*. Georgetown University Press.
- Brandom B (2008) Pragmatism, phenomenalism, and truth talk. *Midwest Studies In Philosophy* 12, 75–93. <https://doi.org/10.1111/j.1475-4975.1988.tb00159.x>.
- Bridge G (2008) City senses: On the radical possibilities of pragmatism in geography. *Geoforum* 39(4), 1570–1584. <https://doi.org/10.1016/j.geoforum.2007.02.004>.
- Bridgman P and Davis G (2003) What use is a policy cycle? Plenty, if the aim is clear. *Australian Journal of Public Administration* 62(3), 98–102. <https://doi.org/10.1046/j.1467-8500.2003.00342.x>.
- Buchler J (1955) *Philosophical Writings of Peirce*. Dover Publications, Inc.
- Capps J (2018) Did Dewey have a theory of truth? *Transactions of the Charles S. Peirce Society* 54(1), 39–63. <https://doi.org/10.2979/trancharpeirsoc.54.1.03>.
- Capps J (2019) The pragmatic theory of truth. In Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2019/entries/ruth-pragmatic/>.
- Chailleux S and Zittoun P (2022) A pragmatist constructivism framework (PCF) to understand the undetermined policy process: Three policy solutions for shale gas problem. *Policy Studies* 44, 276–296. <https://doi.org/10.1080/01442872.2021.2024160>.
- Chaudhari S, Aggarwal P, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K, Deshpande A and da Silva BC (2024) *RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs*. arXiv. arXiv:2404.08555.
- Cherryholmes CH (1994) More notes on pragmatism. *Educational Researcher* 23(1), 16–18. <https://doi.org/10.2307/1176282>.
- Conitzer V, Freedman R, Heitzig J, Holliday WH, Jacobs BM, Lambert N, Mossé M, Pacuit E, Russell S, Schoelkopf H, Tewolde E and Zwicker WS (2024) *Social Choice for AI Alignment: Dealing with Diverse Human Feedback*. arXiv. arXiv:2404.10271.
- Cormier H (2000) *The Truth Is What Works: William James, Pragmatism, and the Seed of Death*. Rowman & Littlefield Publishers.
- de Freitas Araujo S (2020) Truth, half-truth, and post-truth: Lessons from William James. *Journal of Constructivist Psychology* 35, 478–490. <https://doi.org/10.1080/10720537.2020.1727390>.
- Dewey J (2018) *Logic—The Theory of Inquiry*. Read Books, Ltd.
- Dicker G (1971) John Dewey: Instrumentalism in social action. *Transactions of the Charles S. Peirce Society* 7(4), 221–232.
- Dunn WN (2018a) Rediscovering pragmatism and the policy sciences. *European Policy Analysis* 4(1), 13–22. <https://doi.org/10.1002/epa2.1038>.
- Dunn WN (2018b, February 26) *Harold Lasswell and the Study of Public Policy*. Oxford Research Encyclopedia of Politics. <https://doi.org/10.1093/acrefore/9780190228637.013.600>.
- Dunn WN (2019) *Rediscovering Pragmatism as the Origin of the Policy Sciences*. The Annual Conference of the International Public Policy Association, Canada.

- Farber DA** (2015) Coping with uncertainty: Cost-benefit analysis, the precautionary principle, and climate change. *Washington Law Review* 90, 1659.
- Fischer F and Miller GJ** (2006) *Handbook of Public Policy Analysis: Theory, Politics, and Methods*. CRC Press.
- Forstener J** (2018) *Something Has Cracked: Post-Truth Politics and Richard Rorty's Postmodernist Bourgeois Liberalism*.
- Garrison J** (1999) John Dewey's theory of practical reasoning. *Educational Philosophy and Theory* 31(3), 291–312. <https://doi.org/10.1111/j.1469-5812.1999.tb00467.x>.
- Goo JJ and Heo, J.-Y.** (2020). The impact of the regulatory sandbox on the Fintech industry, with a discussion on the relation between regulatory sandboxes and open innovation. *Journal of Open Innovation: Technology, Market, and Complexity*, 6(2), Article 2. <https://doi.org/10.3390/joitmc6020043>.
- Hildebrand D** (2021) John Dewey. In Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2021/entries/dewey/>.
- Hildebrand DL** (2005) Pragmatism, neopragmatism, and public administration. *Administration & Society* 37(3), 345–359. <https://doi.org/10.1177/0095399705276114>.
- Hookway C** (2008) The pragmatist maxim and the proof of pragmatism (2) after 1903. *Cognition* 9(1), 57–72.
- Hookway C** (2009) Habits and Interpretation: Defending the Pragmatist Maxim. 14.
- Hookway C** (2012) *The Pragmatic Maxim: Essays on Peirce and Pragmatism*. OUP Oxford.
- Huang LT-L, Papsyshev G and Wong JK** (2024) Democratizing value alignment: From authoritarian to democratic AI ethics. *AI and Ethics*. <https://doi.org/10.1007/s43681-024-00624-1>.
- Iyasu D** (2009) Ethiopia's salvation: Pragmatism in public health policy-making. *Journal of Public Health Policy* 30(3), 280–284. <https://doi.org/10.1057/jphp.2009.27>.
- James W** (1975) *Pragmatism (1–1)*. Harvard University Press.
- Kaufmann T, Weng P, Bengs V and Hüllermeier E** (2023) A survey of reinforcement learning from human feedback. arXiv. arXiv:2312.14925.
- Kaufman-Osborn TV** (1985) *Pragmatism, Policy Science, and the State*. <https://doi.org/10.2307/2111183>.
- Kelly LM and Cordeiro M** (2020) Three principles of pragmatism for research on organizational processes. *Methodological Innovations* 13(2), 2059799120937242. <https://doi.org/10.1177/2059799120937242>.
- Knight J and Johnson J** (1999) Inquiry into democracy: What might a pragmatist make of rational choice theories? *American Journal of Political Science* 43(2), 566–589. <https://doi.org/10.2307/2991807>.
- Krakovna V, Orseau L, Ngo R, Martic M and Legg S** (2020) Avoiding side effects by considering future tasks. *Advances in Neural Information Processing Systems* 33, 19064–19074.
- Kumar C** (2005) Foucault and Rorty on truth and ideology: A pragmatist view from the left. *Contemporary Pragmatism* 2(1), 35–94. <https://doi.org/10.1163/18758185-90000003>.
- Kundu S, Bai Y, Kadavath S, Askill A, Callahan A, Chen A, Goldie A, Balwit A, Mirhoseini A, McLean B, Olsson C, Evraets C, Tran-Johnson E, Durmus E, Perez E, Kernion J, Kerr J, Ndousse K, Nguyen K, ... Kaplan J** (2023). Specific Versus General Principles for Constitutional AI. arXiv. arXiv:2310.13798.
- Lasswell HD** (1956) *The Decision Process: Seven Categories of Functional Analysis*. Bureau of Governmental Research, College of Business and Public Administration, University of Maryland.
- Lasswell HD** (1970) The emerging conception of the policy sciences. *Policy Sciences* 1(1), 3–14.
- Lee FLF** (2010) Pragmatism, perceived reality, and Hong Kong People's attitudes toward democratic reform. *Issues & Studies* 46(1), 189–219.
- Lee K, Hwang D, Park S, Jang Y and Lee M** (2024) Reinforcement Learning from Reflective Feedback (RLRF): Aligning and Improving LLMs via Fine-Grained Self-Reflection. arXiv. arXiv:2403.14238.
- Legg C and Hookway C** (2021) Pragmatism. In Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2021/entries/pragmatism/>.
- Li A, Xiao Q, Cao P, Tang J, Yuan Y, Zhao Z, Chen X, Zhang L, Li X, Yang K, Guo W, Gan Y, Yu X, Wang D and Shan Y** (2024) HRLAIF: Improvements in Helpfulness and Harmlessness in Open-domain Reinforcement Learning From AI Feedback. arXiv. arXiv:2403.08309.
- Lyons D** (1981) Legal formalism and instrumentalism—a pathological study. *Cornell Law Review* 66(5), 949–972.
- Malkan J** (1997) Literary formalism, legal formalism. *Cardozo Law Review* 19, 1393.
- Marchetti, S.** (2010). William James on truth and invention in morality. *European Journal of Pragmatism and American Philosophy* II(2), Article 2. <https://doi.org/10.4000/ejppap.910>.
- Miller H** (2004) Why Old Pragmatism Needs an Upgrade: [1]—Proquest. <https://www.proquest.com/docview/196832775?pq-origsite=gscholar&fromopenview=true>.
- Minteer BA, Corley EA and Manning RE** (2004) Environmental ethics beyond principle? The case for a pragmatic contextualism. *Journal of Agricultural & Environmental Ethics* 17(2), 131–156. <https://doi.org/10.1023/B:JAGE.0000017392.71870.1f>.
- Nair S and Howlett M** (2016) Meaning and power in the design and development of policy experiments. *Futures* 76, 67–74. <https://doi.org/10.1016/j.futures.2015.02.008>.
- Newman JO** (1984) Between legal realism and neutral principles: The legitimacy of institutional values. *California Law Review* 72(2), 200–216. <https://doi.org/10.2307/3480475>.
- Oldenhof L, Wehrens R and Bal R** (2022) Dealing with conflicting values in policy experiments: A new pragmatist approach. *Administration & Society*, 00953997211069326. <https://doi.org/10.1177/00953997211069326>.

- Olshewsky TM** (1983) Peirce's pragmatic maxim. *Transactions of the Charles S. Peirce Society* 19(2), 199–210.
- Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J and Lowe R** (2022) Training language models to follow instructions with human feedback. arXiv. arXiv:2203.02155.
- Overdeest C, Bleicher A and Gross M** (2010) The experimental turn in environmental sociology: Pragmatism and new forms of governance. In Gross M and Heinrichs H (eds.), *Environmental Sociology: European Perspectives and Interdisciplinary Challenges*. Netherlands: Springer, pp. 279–294. https://doi.org/10.1007/978-90-481-8730-0_16.
- Peters CJ** (2014) *Legal Formalism, Procedural Principles, and Judicial Constraint in American Adjudication* (SSRN Scholarly Paper 2573912). Social Science Research Network. <https://papers.ssrn.com/abstract=2573912>.
- Pihlström S** (2008) How (not) to write the history of pragmatist philosophy of science? *Perspectives on Science* 16(1), 26–69. <https://doi.org/10.1162/posc.2008.16.1.26>.
- Pildes RH** (1999) Forms of formalism. *The University of Chicago Law Review* 66(3), 607–621. <https://doi.org/10.2307/1600419>.
- Posner RA** (1986) Legal formalism, legal realism, and the interpretation of statutes and the constitution. *Case Western Reserve Law Review* 37, 179.
- Pouryousefi S and Freeman RE** (2021) The promise of pragmatism: Richard Rorty and business ethics. *Business Ethics Quarterly* 31(4), 572–599. <https://doi.org/10.1017/beq.2021.6>.
- Prasad, M.** (2018). Social choice and the value alignment problem. In *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC.
- Putnam H** (1995) Pragmatism. *Proceedings of the Aristotelian Society* 95, 291–306.
- Putnam RA** (2006) William James and moral objectivity. *William James Studies*, 1. <https://www.jstor.org/stable/26203682>.
- Radin M and Wagner RP** (1999) The myth of private ordering: Rediscovering legal realism in cyberspace. *Chicago-Kent Law Review*. https://scholarship.law.upenn.edu/faculty_scholarship/744.
- Ramberg B and Dieleman S** (2021) Richard Rorty. In Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2021/entries/rorty/>.
- Rex J** (2020) Anatomy of agency capture: An organizational typology for diagnosing and remedying capture. *Regulation & Governance* 14(2), 271–294. <https://doi.org/10.1111/rego.12209>.
- Rigby J** (2005) Policy, theory and pragmatism: Implementing the UK's energy efficiency best practice Programme. *Policy & Politics* 33(2), 277–296. <https://doi.org/10.1332/0305573053870149>.
- Ringe W-G and Ruof C** (2018) *A Regulatory Sandbox for Robo Advice* (SSRN Scholarly Paper 3188828). Social Science Research Network. <https://doi.org/10.2139/ssrn.3188828>.
- Rorty R** (1995) Is truth a goal of enquiry? Davidson vs Wright. *The Philosophical Quarterly* 45(180), 281–300. <https://doi.org/10.2307/2219651>.
- Rudolph JL** (2005) Inquiry, instrumentalism, and the public understanding of science. *Science Education* 89(5), 803–821.
- Sager F** (2007) Habermas' models of decisionism, technocracy and pragmatism in times of governance: The relationship of public administration, politics and science in the alcohol prevention policies of the Swiss member states. *Public Administration* 85(2), 429–447. <https://doi.org/10.1111/j.1467-9299.2007.00646.x>.
- Schauer F** (2012) *Legal Realism Untamed* (SSRN Scholarly Paper 2064837). Social Science Research Network. <https://doi.org/10.2139/ssrn.2064837>.
- Shields PM** (1996) Pragmatism: Exploring public administration's policy imprint. *Administration & Society* 28(3), 390–411. <https://doi.org/10.1177/009539979602800305>.
- Shields PM** (2003) The community of inquiry: Classical pragmatism and public administration. *Administration & Society* 35(5), 510–538. <https://doi.org/10.1177/0095399703256160>.
- Shields PM** (2008) Rediscovering the taproot: Is classical pragmatism the route to renew public administration? *Public Administration Review* 68(2), 205–221. <https://doi.org/10.1111/j.1540-6210.2007.00856.x>.
- Shuklian S** (1995) Marx, Dewey, and the instrumentalist approach to political economy. *Journal of Economic Issues* 29(3), 781–805.
- Sidorsky D and Talisse R** (2018) Sidney Hook. In Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2018/entries/sidney-hook/>.
- Singer JW** (1988) Legal realism now. *California Law Review* 76, 465.
- Sosa E** (1993) Putnam's pragmatic realism. *The Journal of Philosophy* 90(12), 605–626. <https://doi.org/10.2307/2940814>.
- Stone D** (1998) Principles and pragmatism in the privatisation of British higher education. *Policy and Politics* 26(3), 255–271. <https://doi.org/10.1332/030557398782213674>.
- Tamanaha BZ** (2008) *Understanding Legal Realism* (SSRN Scholarly Paper 1127178). Social Science Research Network. <https://doi.org/10.2139/ssrn.1127178>.
- Thayer HS** (1977) On William James on truth. *Transactions of the Charles S. Peirce Society* 13(1), 3–19.
- The OECD Artificial Intelligence Policy Observatory** (n.d.) Retrieved January 9, 2021, from <https://www.oecd.ai/>.
- Tjalve VS** (2013) Realism, pragmatism and the public sphere: Restraining foreign policy in an age of mass politics. *International Politics* 50(6), 784–797. <https://doi.org/10.1057/ip.2013.41>.
- Troop P** (2018) Why legal formalism is not a stupid thing. *Ratio Juris* 31(4), 428–443. <https://doi.org/10.1111/raju.12225>.
- Truby J, Brown RD, Ibrahim IA and Parellada OC** (2021) A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation* 13, 270–294. <https://doi.org/10.1017/err.2021.52>.

- Ulrich W** (2007) Philosophy for professionals: Towards critical pragmatism. *Journal of the Operational Research Society* 58(8), 1109–1113. <https://doi.org/10.1057/palgrave.jors.2602336>.
- Unger RM** (1983) The critical legal studies movement. *Harvard Law Review* 96(3), 561–675. <https://doi.org/10.2307/1341032>.
- Unger RM** (2021) *The Universal History of Legal Thought*. Deep Freedom Books.
- Upin JS** (1993) Charlotte Perkins Gilman: Instrumentalism beyond Dewey. *Hypatia* 8(2), 38–63.
- Weinrib EJ** (1993) The jurisprudence of legal formalism. *Harvard Journal of Law & Public Policy* 16(3), 583.
- Welchman J** (1989) From absolute idealism to instrumentalism: The problem of Dewey's early philosophy. *Transactions of the Charles S. Peirce Society* 25(4), 407–419.
- West K and Davis P** (2010) What is the public value of government action? Towards a (new) pragmatic approach to values questions in public endeavours. *Public Administration* 89, 226–241. <https://doi.org/10.1111/j.1467-9299.2010.01847.x>.
- White P** (Director) (2008) *Rorty on Truth* [Video recording]. <https://www.youtube.com/watch?v=CzynRPP9XkY>.
- Wible JR** (1984) The instrumentalisms of Dewey and Friedman. *Journal of Economic Issues* 18(4), 1049–1070.
- Williams M** (2003) Rorty on knowledge and truth. In *Richard Rorty* (pp. 61–80). Cambridge university press.
- Williamson OE** (1996) Revisiting legal realism: The law, economics, and organization perspective. *Industrial and Corporate Change* 5(2), 383–420. <https://doi.org/10.1093/icc/5.2.383>.
- Züger T and Asghari H** (2023) AI for the public. How public interest theory shifts the discourse on AI. *AI & Society* 38(2), 815–828. <https://doi.org/10.1007/s00146-022-01480-5>.