Short Paper: Causal Anticipation for Reason-Based AI Alignment

Yannic Muskalla
DFKI
Saarbrücken, Germany
yannic.muskalla@dfki.de

Recent proposals for the normative alignment of AI agents advocate grounding morally relevant decisions in normative reasons, formalized using symbolic rules [4, 1]. These rules can be embedded into agents in a way that makes their compliance verifiable. I call this approach *reason alignment*. Unlike classical outcome alignment methods, which aim to ensure that an agent does the right thing, reason alignment aspires to ensure that the agent does the right thing, in permissible ways, for the right reasons. However, it is not feasible to enter every reason for action into a reason theory by hand. Therefore, this paper takes a first step into automatically deducing reasons for action. For this purpose, I introduce *Causal Normative Models*, a framework that extends standard causal models with normative rules, whose antecedents contain events formulated over causal variables. Within this framework, I identify sufficient conditions under which given normative rules entail additional *precursory rules*, corresponding to causal precursors of the original reasons.

1 Introduction

Alignment Classical AI alignment methods suffer from fundamental problems. On the one hand, they lack justification and thus authors call for a broader public justificatory basis [7]. On the other hand, their robustness is frequently questionable. Methods related to Reinforcement Learning from Human Feedback (RLHF) intransparently combine instrumental demands as well as normative requirements into one single reward signal. This method is not able to provide safety guarantees [5] and opens the possiblilty for reward hacking. On the other side, there are alignment methods that take a hardcoded list of deontological constraints (e.g. Anthropic's Constitutional AI method [3]), which might be able to produce safety guarantees, but are not able to contextually adapt during runtime.

Several authors have recently proposed grounding AI alignment in reasons [1, 4]. I refer to such approaches as *reason alignment*. This idea parallels the *reasons-first* approach in (meta)ethics, which seeks to ground the deontic status (e.g. the moral status) of an action in the normative reasons that count in its favor [12, 6, 13]. Analogously, reason alignment aims to constrain an AI agent by providing it with a set of normative reasons for action, which the agent is required to follow whenever those reasons are pertinent. This paper does not aim to defend reason alignment against competing approaches, but to show how it can be extended in a specific way to enhance its practical applicability.

Reasons In this article, I focus on *normative justifying reasons for actions*, which I will, for the sake of brevity, refer to simply as *reasons*. A reason is something that speaks in favor of acting in a particular way [2]. While there is an ongoing debate about the ontology of reasons [2], I will follow the mainstream and take reasons to be *facts*, understood as states of affairs that actually obtain. This enables a useful

Submitted to:
FEAR 2025

distinction: A fact that speaks in favor of an action φ is an *actual reason* for φ, whereas a state of affairs (or in short: state) is a *potential reason* for φ if it would count as an actual reason were it to obtain. Unless noted otherwise, I use 'reasons' to mean potential reasons.

Aim of this paper There is an intuitive connection between preventing and ending undesirable states. If a state provides reason to end it once it obtains, then we also have reason to act preventively, to keep it from arising in the first place. To phrase it in terms of duties: if one has a duty to end something, one likewise has an anticipatory duty to prevent its occurrence. For example, if there is a moral duty to mitigate or remedy a certain harm, then there is likewise a moral duty to prevent that harm from occurring. If we have a duty to rescue drowning persons, we also have a duty to prevent individuals staggering near a river from falling into the water. Of course, such anticipatory reasons (or duties) may be outweighed or defeated by competing ones. The point here is merely that these anticipatory reasons exist, not that they necessarily generate an all-things-considered obligation.

Given the intuitive plausibility, the question of under which conditions such anticipatory reasons and duties exist is an interesting one in ethics.[1] It becomes even more significant in the context of reason alignment: manually specifying every potentially relevant normative reason for action through human oversight will often be infeasible. Hence, methods for deriving anticipatory reasons are highly valuable, as they offer an automated means of extending a given reason theory. The following section introduces the formal machinery needed to formulate sufficient conditions for the derivation of anticipatory reasons.

2 Towards Causal Normative Models

2.1 Causal Models

A *causal model* $\mathcal{M}$ is generally understood as a pair $(\mathcal{S},\mathcal{F})$ consisting of a signature $\mathcal{S}$ and a set $\mathcal{F}$ of functions. The signature $\mathcal{S} = (U,V,R)$ consists of

- a set of exogenous variables U,
- a set of endogenous variables V, and
- a range R, that assigns each variable in $U \cup V$ the set of values it can take.

A variable is exogenous, iff its value is not determined by the other variables in the model.

The set of functions $\mathcal{F}$ contains a function for each endogenous variable that describes how the value of this variable causally depends directly on the values of the other variables of the model. It is common to note down these functions as structural equations. So e.g. instead of $f_Y(X,U_i) = 2X + U_i$, I write $Y \coloneqq 2X + U_i$.

Such causal models $\mathcal{M} = (\mathcal{S},\mathcal{F})$ can be evaluated with respect to a context $\vec{u}$. A context is a function $U \to R(U)$ that assigns each exogenous variable a value. Note that the values of the exogenous variables uniquely determine the values of the endogenous variables, so a context $\vec{u}$ describes all variable values. We call a variable X taking on a value x an *event* $X = x$. Additionally, we consider interventions $X \to x$ that set the value of the variable X to x, regardless of its structural equation and without effecting the values of any ancestors of X. This allows to evaluate events with respect to a model and a context. We use the following notation

$$(\mathcal{M},\vec{u}) \vDash [X \leftarrow x]Y = y$$

[1] To the best of my knowledge, there is little work on this question.

to formulate that in model $\mathcal{M}$ and context $\vec{u}$, the event $Y = y$ occurs if an intervention that sets X to x is performed.[2]

2.2 Normative Rules

A *normative rule* is a pair (r, φ), where r is a reason for φ. For instance, the traffic light being red can be considered a normative reason for stopping in front of it. A corresponding rule may be $(\text{red light}, \varphi_{stop})$. It can be read as a conditional obligation: If the light is red, you ought to stop.[3]

In the context of causal models, we already defined events as variables taking on values. To allow for a better integration of rules into causal models, we need to bridge the gap between events and states. Therefore, we assume that to each event, there is a corresponding state describing the event.[4] As an example take a variable L with $R(L) = \{red, yellow, green\}$ indicating the color of a traffic light. The state describing the event $L = red$ is a reason for stopping. I introduce the notation $L = red \circ\!\!\rightarrow \varphi_{stop}$ to denote the rule consisting of the action φ_{stop} and the state describing the event $L = red$.

2.3 Causal Normative Models

I now combine the previous considerations into *causal normative models*, which are causal models that incorporate normative rules. A causal normative model $\mathcal{M}_n$ is a triple $(\mathcal{S}_n, \mathcal{F}, \mathcal{R})$, where the *normative signature* $\mathcal{S}_n = (U, V, R, \phi)$ consists of

- a set of exogenous variables U,
- a set of endogenous variables V,
- a range R, that assigns each variable in $U \cup V$ the set of values it can take, and
- a set of actions ϕ.

Additionally to this signature $\mathcal{S}_n$ and the set of functions $\mathcal{F}$, a causal normative model contains a set of rules $\mathcal{R}$. Each rule $\rho \in \mathcal{R}$ has the form $\vec{X} = \vec{x} \circ\!\!\rightarrow \varphi$, where $\vec{X} = \vec{x}$ is a non-empty conjunction of primitive events $X_1 = x_1 \wedge \ldots \wedge X_n = x_n$ and for each conjunct $X_i = x_i$ it holds that $X_i \in V$ and $x_i \in R(X_i)$. Furthermore, the action φ in the consequent of the rule has to be part of the action set ϕ specified in the normative signature ($\varphi \in \phi$).

2.4 Causal Normative Graphs

For the purposes of illustration, it is quite common to depict causal models as *causal graphs*. In such a graph, each endogenous variable is a node that receives incoming edges from all the endogenous variables that feature in its structural equation. For causal normative models, I extend this representation by adding all the actions in ϕ as nodes. The node of an action φ receives incoming edges from all the endogenous variables that feature in the antecedent of a rule the action is part of. To distinguish causal edges from edges indicating rules, I use normal edges for the former and circled edges ($\circ\!\!\rightarrow$) for the latter dependencies. We call such a graph *causal normative graph*.

[2] For a more extensive introduction into the used notation of causal models for particular events, confer to [8]. For a more general introduction into causal models, consider [11].

[3] Note that if there is normative reason for φ, this does not imply that it is right to φ. Reasons may be outweighed or defeated by other reasons.

[4] One may be discontent with the terminology that a state describes an event. Nothing hinges on this formulation, as long as one is willing to grant that there is a corresponding state to every event.

2.5 Example: Dam Break

Suppose a dam upstream is showing signs of structural weakness. If it fails, the resulting flood will endanger the lives of villagers in downstream settlements. The danger of being flooded after the bursting of the dam is a strong reason to rescue the villagers. But arguably, the signs of structural weakness already give reason to reinforce the dam to prevent the bursting in the first place.

Let us model this as follows: we are interested in three events, namely the dam showing signs of structural weakness, the dam bursting and the villagers being in danger of being flooded. We model these events via three binary variables W, B and D that take on value 1 if the corresponding event occurs and 0 otherwise. Whether $W = 1$ occurs depends on factors, which we do not observe. This leads to the structural equation $W \coloneqq U_W$, where U_W is an exogenous variable modeling unobserved factors. The occurrence of $B = 1$ is causally influenced by W as well as by other unobserved factors, so we model the structural equation for B as $B \coloneqq W \wedge U_B$, where U_B models unobserved causal (and non-counfounding) factors. The villagers are in direct danger of flooding if the dam breaks, so we use the structural equation $D \coloneqq B$.

There are two actions relevant, namely to save the villagers (φ_S) and to reinforce the dam (φ_R). In case the villagers are in danger of being flooded, the reason for rescuing the villagers becomes pertinent. This is encoded as the rule $\rho_1 : D = 1 \circ\!\!\rightarrow \varphi_S$. The dam showing signs of structural weakness in turn is a normative reason to reinforce the dam, which is encoded as $\rho_2 = W = 1 \circ\!\!\rightarrow \varphi_R$. In conclusion, we model the given scenario as

- $U = \{U_W, U_B\}$
- $V = \{W, B, D\}$
- $\forall X \in U \cup V. R(X) = \{0, 1\}$
- $\phi = \{\varphi_R, \varphi_V\}$
- $\mathcal{F} = \{f_W, f_B\}$, with $f_W(U_w) = U_W$ and $f_B(W, U_B) = \max(W, U_B)$
- $\mathcal{R} = \{\rho_1, \rho_2\}$, with $\rho_1 : D = 1 \circ\!\!\rightarrow \varphi_S$ and $\rho_2 : W = 1 \circ\!\!\rightarrow \varphi_R$

For the causal normative graph of this model, confer to Figure 1. Although ρ_1 and ρ_2 are two separate rules with two different reasons, it appears that ρ_2 somehow follows from ρ_1: the structural weakness of the bridge gives us reason to repair the dam, because the structural weakness causes the dam to break and thereby endanger the villagers in the future.

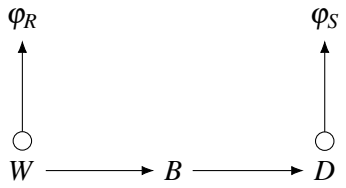

Figure 1: The causal normative graph of the dam break example

3 Causal Anticipatory Reasons

We now have the formal tools to discuss the main question of this article: under which conditions is it the case that normative rules imply causal anticipatory rules?

3.1 Terminology

If an event $C = c$ is a cause of $E = e$, then we call the state describing $C = c$ a causal precursor of the state describing $E = e$.

For the sake of readability and conciseness, we will in the following speak of reasons as if they were events, rather than always referring to the states that describe those events. Accordingly, we will sometimes speak as though reasons themselves stand in causal relations to one another, even though, strictly speaking, it is the events described by the states underlying the reasons that bear these causal relations.

In the framework of causal normative models, we speak of an action φ as *reasonable* if there is an event that actually occurs and is a reason for φ.

3.2 Sufficient Conditions

I will now present three conditions which, in conjunction, are sufficient to identify a causal anticipatory rule from a given rule. Afterwards, we will elaborate on each condition and discuss methods for their operationalization.

Definition 1 (Causal Anticipation) *Given a particular causal normative Model $\mathcal{M}_n$ and two variables $C, E \in V$ and $c \in R(C), e \in R(E)$, as well as two actions $\varphi_C, \varphi_E \in \phi$:*

From $E = e \circ\!\!\rightarrow \varphi_E$, it follows that $C = c \circ\!\!\rightarrow \varphi_C$, if

1. *(Actual Causation) $C = c$ is an actual cause of $E = e$,*
2. *(Effectiveness) performing φ_C in turn prevents $E = e$ from occurring, and*
3. *(Aim at Discontinuation) φ_E aims at preventing $E = e$ from persisting.*

If those condition are jointly met, then we call the state describing $C = c$ an (causally) anticipatory reason and the rule $C = c \circ\!\!\rightarrow \varphi_C$ an anticipatory rule.

3.2.1 Actual Causation

The first condition is that the event appealed to in the anticipatory reason must stand in an actual causal relation to the state that grounds the anticipated reason. This condition forms the basic starting point of the framework.

We require that the events described by the two reasons be causally connected. In particular, the anticipatory reason must describe a cause of the reason invoked in the given rule. If the structural weaknesses were not actually to cause the dam break – for instance, if the dam break were not going to occur at all, or if it were instead the result of a terrorist attack – then no conclusion could be drawn from the fact that the dam break is a reason for saving the villagers to the claim that the structural weakness themselves constitute a reason.

Operationalization As a concrete suggestion on how to spell out this causal dependency, one could help oneself with the definition of actual causation Joseph Halpern gives in his book *Actual Causality* [8, p. 23]:[5]

Definition 2 (Actual Causation) *The non-empty conjunction of primitive events $\vec{X} = \vec{x}$ is an actual cause of the event ε in the causal setting $(\mathcal{M}, \vec{u})$, where $\mathcal{M} = ((U, V, R), \mathcal{F})$, iff*

[5] Note that I present the version of the definition that Halpern calls "modified version".

AC1. $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x}$ *and* $(\mathcal{M}, \vec{u}) \models \varepsilon$.

AC2. There is a subset $\vec{W}$ *of* V *and a set of variable values* $\vec{x}'$ *of the variables in* $\vec{X}$ *such that* $(\mathcal{M}, \vec{u}) \models \vec{W} = \vec{w}$ *and*

$$(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varepsilon.$$

AC3. $\vec{X}$ *is minimal; there is no* $\vec{X}' \subset \vec{X}$*, such that* $\vec{X}' = \vec{x}'$ *satisfies conditions AC1 and AC2, where* $\vec{x}'$ *are those values of* $\vec{x}$*, that belong to the variables in* $\vec{X}'$.

Assuming, in the dam break example, that the dam break actually occurs in $\vec{u}$ if not intervened, we get that $W = 1$ is an actual cause of $D = 1$: AC1 simply demands that both $W = 1$ and $D = 1$ occur. For AC2, we choose $\vec{W} = \varnothing$ and get $(\mathcal{M}_n, \vec{u}) \models [W \leftarrow 0]\neg(D = 1)$. $W = 1$ is not a conjunction, so AC3 trivially holds.

3.2.2 Effectiveness

For a rule $E = e \circ\!\!\rightarrow \varphi_E$ to imply another rule $C = c \circ\!\!\rightarrow \varphi_C$, it does not suffice that $C = c$ causes $E = e$. To make sense of the idea of anticipatory reasons, the action φ_C must also be effective in preventing the otherwise caused event $E = e$.

If we omit this condition then the cause would be a reason for arbitrary actions, e.g. the structural weakness of the dam would count as a reason to go fishing, even though fishing does not help in preventing the dam break. With this condition in place, only actions that are apt to prevent the otherwise caused effect count on that account as reasonable. After the dam showed signs of structural weakness, repairing it will be effective to prevent the otherwise caused danger of flooding for the villagers. Therefore, the structural weakness passes this condition to be a reason to repair the dam.

Operationalization To formalize the requirement stated in the second condition, we extend the model with an additional binary variable, which takes the value 1 if the corresponding action is performed and 0 otherwise. If, in a context where both the actual cause occurs and the action is performed, the actual effect does not occur, then the second condition is satisfied. One can formalize it like this:

Definition 3 (Effectiveness) *Let* $C = c \circ\!\!\rightarrow \varphi_C$ *be a rule. We say that an action* φ_C *is* effective *with respect to an event* $E = e$ *in a model* $\mathcal{M}_n$ *and an actual context* $\vec{u}$ *if, in an adequate adjusted model* $\mathcal{M}_n^A$*, obtained by introducing a new binary variable* A *that takes the value* 1 *iff* φ_C *is performed, it holds that*

$$(\mathcal{M}_n^A, \vec{u}) \models [C \leftarrow c, A \leftarrow 1]\neg(E = e).$$

In the dam break-case, one could introduce an additional binary variable R, that is 1 iff the dam gets repaired. Confer to Figure 2 for the causal normative graph of the adjusted model. Repairing the dam plausibly prevents it from breaking, keeping the villagers safe from flooding. Hence $(\mathcal{M}_n^R, \vec{u}) \models [W \leftarrow 1, R \leftarrow 1]\neg(D = 1)$ holds.

3.2.3 Aim at the Discontinuation

The third and final condition is that φ must aim at preventing the continued persistence of the state describing $E = e$. This condition is admittedly somewhat handwavy, but the underlying idea is straightforward. If φ is not directed at counteracting the persistence of $E = e$, then there is no basis for anticipating and preventing the initial occurrence of $E = e$.

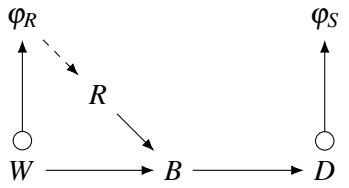

Figure 2: The causal normative graph of the dam break example with intervention variable R. The dashed arrow indicates that R models whether φ_R is performed.

To see why this condition is necessary, consider the following example. An autonomous car might be governed by the normative rule 'children nearby $\circ\!\!\rightarrow$ slow down'. Here, the action of slowing down is clearly not intended to prevent the persistence of its reason. Slowing down does not make the children leave the vicinity. Consequently, no anticipatory rule can be inferred from this case. Without the third condition, however, the above rule would wrongly generate anticipatory rules instructing the car to prevent children from ever approaching in the first place. This is clearly not correct.

Operationalization This condition can be operationalized in a similar fashion as the last condition. We introduce an additional binary variable B that is 1, iff φ_E is performed. We then check whether $B = 1$ prevent the persistence of $E = e$. We therefore introduce another new variable E_2 that displays the value of E at a later timestep and check whether the $B = 1$ leads to $\neg(E_2 = e)$.

Definition 4 (Discontinuation) *Let $E = e \circ\!\!\rightarrow \varphi_E$ be a rule. We say that an action φ_E discontinues an event $E = e$ in a model $\mathcal{M}_n$ and an actual context $\vec{u}$ if, in an adequate adjusted model $\mathcal{M}_n^{A,E_2}$, obtained by introducing a new binary variable B that takes the value 1 iff φ_E is performed and a new variable E_2 that tracks the value of E to a later point in time, it holds that*

$$(\mathcal{M}_n^{A,E_2}, \vec{u}) \models [E \leftarrow e, B \leftarrow 1]\neg(E_2 = e).$$

In the dam break example, this can be applied as follows: we extend $\mathcal{M}_n$ by a binary variable S that is 1 iff the agent saves the villagers from the flood and by another variable D_2 that tracks whether the villagers are still in danger after the saving efforts of the agent. After being saved, the villagers are plausibly out of danger. Therefore, $(\mathcal{M}_n^{S,D_2}, \vec{u}) \models [D \leftarrow 1, S \leftarrow 1]\neg(D_2 = 1)$ holds, φ_S discontinues D_2 and hence the third condition is also satisfied in the example. Confer to Figure 3 for a depiction of the relevant part of the graph of the adjusted model.

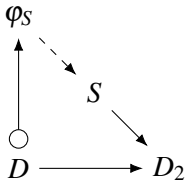

Figure 3: Part of the causal normative graph of the villager example, adjusted to check the discontinuation condition.

3.3 Path to Enhanced Reason Alignment

As discussed, the three conditions are sufficient to determine, with respect to a given rule, whether a candidate rule qualifies as an anticipatory rule. They can therefore be used to extend a set of rules $\mathcal{R}$ (a reason theory) by adding all anticipatory rules that can be derived through the application of these three conditions. That is, we can construct the extension $\mathcal{R}^a$, such that $\mathcal{R}^a$ contains all rules of $\mathcal{R}$ as well as all rules ρ_i for which there exists a rule $\rho_j \in \mathcal{R}$ such that ρ_i is an anticipatory rule with respect to ρ_j.

Reason alignment architectures can employ this procedure to ensure that an aligned AI system acts in anticipation. Consider the dam-break case: if the reason theory of the system only contains the rule $\rho_1 : D = 1 \circ\!\!\rightarrow \varphi_S$, the alignment mechanism becomes effective only once the dam has already failed and the villagers are in danger of flooding. To align the AI such that it acts preventively, a corresponding rule such as ρ_2 would normally have to be added manually. However, this is no longer necessary under the approach developed in this paper. By extending the reason theory through its closure with respect to Definition 1, the anticipatory rule ρ_2 is automatically derived, compelling the AI to act in anticipation of the dam break.

It thus suffices that the reason theory contains at least one rule at the end of a causal chain. The corresponding anticipatory rules can then be derived automatically. This moves toward the broader vision of *reason-based AI alignment*: a framework centered on a reason theory that integrates both human-provided rules and derived, anticipatory rules learned from them.

4 Conclusion & Future Work

In this paper, I introduced *causal normative models*, a framework for incorporating normative rules into causal models. This framework made it possible to articulate three conditions that are jointly sufficient to infer an anticipatory rule from a given rule. The three conditions, *Actual Causation*, *Effectiveness*, and *Aim at Discontinuation*, admit a clear operationalization. This, of course, presupposes the availability of a causal discovery algorithm that correctly identifies the underlying causal model,[6] which can then be extended with rules and action variables.

As a starting point, I developed the three conditions within deterministic causal models. To make these considerations practically applicable for aligning agents, one must take (epistemic) uncertainty into account and thus work with a probabilistic notion of causality and with probabilistic causal models. Future work should therefore examine how the proposed conditions can be extended to the probabilistic setting. Another important direction is to explore richer conceptions of reasons. For example, in Nair and Horty's logic of reasons [10], reasons are partially ordered and may be defeated, while in [1] reasons are assigned weights. How anticipatory reasons integrate into such frameworks remains an open question.

References

[1] Benoît Alcaraz, Aleks Knoks & David Streit (2024): *Estimating Weights of Reasons Using Metaheuristics: A Hybrid Approach to Machine Ethics*. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7(1), pp. 27–38, doi:10.1609/aies.v7i1.31614. Available at `https://ojs.aaai.org/index.php/AIES/article/view/31614`.

[6] For the *Action Causation* condition, however, it is not sufficient to recover the causal graph; the structural equations themselves are also required.

[2] Maria Alvarez & Jonathan Way (2024): *Reasons for Action: Justification, Motivation, Explanation*. In Edward N. Zalta & Uri Nodelman, editors: *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition, Metaphysics Research Lab, Stanford University.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown & Jared Kaplan (2022): *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073.

[4] Kevin Baum, Lisa Dargasz, Felix Jahn, Timo P. Gros & Verena Wolf (2024): *Acting for the Right Reasons: Creating Reason-Sensitive Artificial Moral Agents*. arXiv:2409.15014.

[5] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh & Dylan Hadfield-Menell (2023): *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. arXiv:2307.15217.

[6] Jonathan Dancy (2004): *Ethics Without Principles*. Oxford University Press, doi:10.1093/0199270023.001.0001. Available at `https://doi.org/10.1093/0199270023.001.0001`.

[7] Iason Gabriel & Geoff Keeling (2025): *A matter of principle? AI alignment as the fair treatment of claims*. *Philosophical Studies* 182, pp. 1951–1973, doi:10.1007/s11098-025-02300-4.

[8] Joseph Y. Halpern (2016): *Actual Causality*. The MIT Press. Available at `http://www.jstor.org/stable/j.ctt1f5g5p9`.

[9] David Lewis (1973): *Causation*. *Journal of Philosophy* 70(17), pp. 556–567, doi:10.2307/2025310.

[10] Shyam Nair & John Horty (2018): *The Logic of Reasons*. In Daniel Star, editor: *The Oxford Handbook of Reasons and Normativity*, Oxford University Press, pp. 67–84.

[11] Judea Pearl (2000): *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York.

[12] Joseph Raz (1999): *Practical Reason and Norms*. Oxford University Press, doi:10.1093/acprof:oso/9780198268345.001.0001. Available at `https://doi.org/10.1093/acprof:oso/9780198268345.001.0001`.

[13] Mark Schroeder (2021): *Reasons First*. Oxford University Press.