Sentientism, Decision Theory, and Moral Worlds

Frank P. DeVita

May 2026

1. *Introduction*

It is a responsibility of moral philosophy to respond to the many kinds of ethical situations there might be. However, sometimes an ethical stance isn't conceptually equipped to adequately address a choice situation. If an agent's moral stance is to not harm other people, how do they reason about affecting wildlife, rivers and oceans, or the planet Earth? Should they change their stance, or follow the prescriptions of their chosen moral code? To reason coherently and decide to act, I think an agent can *flex* to different moral codes as choice situations demand them, since some decision contexts will inevitably outrun our default moral commitments. This flexibility is a form of moral intelligence.

In this paper, I consider ways in which *sentientism*—belief in the view that the capacity for conscious positively or negatively charged experiences determines whether an entity deserves moral consideration—can fail to guide action, and develop a flexible theory of moral decision making called *focal reasoning* to navigate what I call *moral worlds* to model and evaluate the permissibility of actions under different ethical

frameworks. With moral worlds and focal reasoning, moral agents are freed from the bounds of any single ethical framework, creating a movable moral center of gravity for decision making in different choice situations. The approach also enables different ways to reason about the *moral circle*—the conceptual boundary that defines which entities are owed moral consideration.

First, I consider the merits and drawbacks of sentientism, and highlight edge cases that I think motivate the need for a flexible picture of moral reasoning. I then test the dynamics of sentientist reasoning with decision theory to expose some weak points. To cover the gaps, I develop the focal reasoning with moral worlds to think about moral reasoning and choice at the level of cognitive attitudes, and consider some objections. The upshot is that the moral worlds approach can guide action and serve as a model for moral thought across different choice situations in a coarse-grained way that complements more formal decision theoretic analyses of action.

2. *The Edge of Sentientism*

How should we reason about the morality of our actions toward different entities? Will one factor or theory suffice? Some philosophers argue that *sentientism*—the view that consciousnesses or sentience ground moral status—might (Birch 2024). *Consciousness* means the capacity to have first person subjective experiences, *sentience*

means consciousness *plus* the capacity to have positively or negatively "charged" or *valenced* experiences, such as pleasure and pain, and *moral status* is the degree to which an entity deserves moral consideration in ethical decision making (Chalmers forthcoming, Clatterbuck and Fischer 2025, Birch 2024). Sentientism grants moral status to entities based on their capacity for valenced experiences, thus admitting them to the *moral circle*—the class of entities with moral status (Sebo 2025). The decision strategy that follows from sentientism is clear: when deciding to act, avoid actions that cause pain or negatively valenced experiences for entities capable of having them.

However, determining whether an entity is sentient isn't straightforward. As humans, we can use others' testimony and genetic similarity to infer that they have experiences like ours. In knowing that others have capacities for perception, reason, language, emotion, and consciousness, we also know they share the capacity to *suffer*—to experience highly unpleasant emotional states associated with considerable pain or distress (DeGrazia 1998). However, this inference is less smooth for non-human entities. Since we don't know what it's like to be a bat (Nagel 1973) or any other creature, *grounding* the moral status of non-human entities is difficult. In decision contexts involving non-humans, the central moral question is, "not whether these entities can think, reason, or speak, but whether they also suffer" (Betham 1789).

Detecting whether an entity has the capacity for phenomenal consciousness is notoriously difficult, and an active area of active philosophical and scientific

investigation in science in the cases of infants (Frolich 2024, Passos-Ferreira 2024), animals (Andrews et al. 2025, Birch et al. 2022), AI systems (Butlin et al. 2025), and beyond humans generally (LeDoux et al. 2023). It isn't clear how exactly to detect sentience or consciousness systematically, there is disagreement about the appropriate operative theory of consciousness, and there is lack of consensus on what kinds of physiological or non-verbal markers are relevant across species (Pennartz et al. 2019).

The uncertainty inherent in detecting sentience generates moral and epistemic uncertainty challenges for sentientist reasoning. One puzzle case for sentientism is a conscious entity with *no* valenced mental states, or a "philosophical Vulcan" who is conscious, but has only neutral experiences (Chalmers forthcoming, Chalmers 2022). It would still be wrong to harm a Vulcan because they would witness and know something bad was happening to them. Sentientism misses this case because it focuses on *valenced* experience. A broader view like *phenomenocentrism* that links the capacity for phenomenal experience *simpliciter* to moral status—regardless of its valence—would grant moral consideration to Vulcans, and result in better decision making should one encounter such an entity.

A sentientist might contend that Vulcans would have valenced experiences in virtue of having interests. However, valenced experience and interests can come apart. A Vulcan might have intellectual or procedural interests in writing, watching films, protecting their family, or doing research, but feel entirely *flat* about these actions and

duties in performance and reflection. Valence doesn't *necessarily* enter the picture, nor is it necessarily required to explain the Vulcan's behavior. Empirical evidence also suggests that goal-directed survival behaviors that *look* valence-driven can in fact be driven by satiety, metabolism, or otherwise non-valenced mechanisms (De Araujo 2008).[1] Therefore, phenomenal consciousness or "raw feel"—understood as *broader* than sentience and *not* entailing valence—looks better for reasoning about the moral circle. Nevertheless, this view is also haunted by the empirical difficulties above.

3. *Sentientism and Decision Theory*

How do we decide under potentially high degrees of uncertainty about sentience or consciousness? *Decision theory* can quantify uncertainty and the permissibility of actions in different choice situations, and represent different risk attitudes agents might have given uncertainty about sentience. A quantitative method is to incorporate representations of uncertainty and risk attitudes into decision theorems (Sebo 2023 and Buchak 2019, respectively). On this approach, *weights* are given to the expected utilities of different outcomes of an action as a function of an agent's degree of belief, or

[1] In these experiments, mice without the capacity for taste continued to consume sugar water despite being given the option to consume regular water. The philosophical significance of this result is that valenced gustatory experience isn't necessary for nourishment-driven behavior and survival.

credence, that an entity is sentient, and how morally risky that agent thinks it would be to act as if it weren't.

There are a few different ways to think about and formally represent this situation. We might express this notion formally as P(*A* | *S*), a *conditional probability* (Lewis 1976) of action *A*, defined as the likelihood of some decision about how to act given a sentience estimation *S* for an entity. We might also express this as a *probability of the conditional* (Stalnaker 1972) P(*S* □→ A) that some evidence of sentience *S would imply* a decision to act a certain way. I prefer the latter for its counterfactual form and normative content, since we are dealing with attitudes about how the way we think an entity might be—i.e., sentient or not—should determine whether we should take some action. To further to shed light on how the expected utilities of our actions change given probabilities of sentience, risk attitudes, and expected utility of actions, we can formalize these relationships as[2]:

$$\sum {P_S}^a \times U_\phi + \sum {P_{\neg S}}^b \times U_\phi$$

Where P_S and $P_{\neg S}$ are probabilities, subjective credences, or degrees of belief that that an entity is sentient or not, respectively, taking a number ≥ 0 and ≤ 1. U_ϕ is the

[2] The form of this equation is inspired by Buchak's 2019 use of risk attitude exponents, and Barry and Cullity's 2022 use of two sums of expected utilities.

expected utility U of some action ϕ taking any real number. The exponents a and b represent the weights of a sentience estimate based on an agent's *risk attitudes*. Risk averse attitudes are represented by a or $b > 1$ on either a good (high utility) outcome, and a or $b = 1$ on bad (low utility) outcomes because they weigh bad outcomes more heavily, while a *risk tolerant* agent would have the reverse weightings, and a *risk neutral* agent would be represented by $a = 1$ and $b = 1$. Both addends are sums to account for the number of individuals of an entity type, and the whole operation is a sum since we are calculating the utility of one action versus another.

Consider cognitive models of three agents with different risk attitudes and different estimates of sentience for some entity in the context of the generic action HARMING, who are wondering whether the action is morally permissible. Assume the agents are sentientists who assign high utility to sentient entity welfare. We can model these agents using the decision theorem above:

HARMING

Agent	Moral framework	Sentience estimate	Utility if sentient	Utility if not sentient	Sentience-adjusted EU	Risk attitude	Attitude-adjusted EU
1	Sentientism	0.75	-20	10	-12.5	Averse	-14.375
2	Sentientism	0.55	-20	10	-6.5	Tolerant	-1.55
3	Sentientism	0.35	-20	10	-0.5	Neutral	-0.5

How ought these agents to make choices about how to act? Let's stipulate that actions with utility ≤ 0 are permissible and those with utility > 0 are impermissible. If these agents follow their sentience estimates, they should all refrain from HARMING. Here, the role of risk attitudes is significant—if an agent has high sentience estimates, should their attitude carry less weight? Should attitudes carry more weight in low sentience estimate contexts? The difficulty in answering these questions could be reason to deny the permissibility of risk tolerance or risk aversion, however this seems unfaithful to the cognitive phenomenology of choice, in which our attitudes play a crucial role. If we were to eschew risk attitude adjustment in decision theorems, this would impoverish the cognitive modeling of choosers, so we ought to incorporate it.

Let's return to our model. Increasing sentience estimates and adjusting again for risk attitudes, HARMING has decisively negative utility for sentientists:

HARMING – HIGH AND EQUAL SENTIENCE

Agent	**Moral framework**	**Sentience estimate**	**Utility if sentient**	**Utility if not sentient**	**Sentience-adjusted EU**	**Risk attitude**	**Attitude-adjusted EU**
1	Sentientism	0.9	-20	10	-17	Averse	-17.9
2	Sentientism	0.9	-20	10	-17	Tolerant	-15.2
3	Sentientism	0.9	-20	10	-17	Neutral	-17

So far so good. However, things get interesting when sentience estimates are low or zero in a risk averse, precautionary agent—the kind of agent influential sentientist accounts contend we ought to be (Fischer et al. 2025, Sebo 2018, Birch 2017). Consider a context in which entities that have uncertain, low, or zero sentience estimates are involved in the choice situation, e.g. an ecosystem, AI system, forest, river, or the planet Earth. Ramping down sentience estimates yields:

HARMING – LOW OR NO SENTIENCE

Agent	Moral framework	Sentience estimate	Utility if sentient	Utility if not sentient	Sentience-adjusted EU	Risk attitude	Attitude-adjusted EU
1	Sentientism	0.5	-20	10	-5	Averse	-7.5
2	Sentientism	0.2	-20	10	4	Averse	2.4
3	Sentientism	0	-20	10	10	Averse	10

In these cases, a sentientist is moved to assign *positive* utility to HARMING. This seems like the wrong moral result. It seems intuitively wrong to burn down the California redwoods in Muir Woods, dump toxic waste into New York's Hudson River, or pollute the Earth's atmosphere—not only for instrumental reasons related to how those actions will negatively affect sentient entities in or near those ecosystems, but because those

actions would be deleterious for the entities themselves. Sentientism, although instructive in many situations, is inert in these morally charged choice contexts.

Sentientist reasoning also has a *threshold problem* when viewed decision theoretically. It generates a sorites paradox when one attempts defining an inflection point above which moral status is granted and below which it is not. 99% confidence in the sentience of an entity means moral circle inclusion. The same goes for 98%, 97%, 90%, probably 80% and maybe even 70%, but what about lower levels, parity, or below? Even below 30%, *some* confidence in sentience means that an entity *might in fact* be conscious, and therefore poses a moral hazard if harmed, and moral damage may rapidly proliferate as a function of how many individuals are being affected by an action. Some sentience-based threshold of moral significance is needed to reason within sentientism, but it is unclear how to define it.

It is philosophically unsatisfying for sentience estimation to be the sole mechanism for reasoning about the moral circle. Sentience estimates are uninformative in cases where sentience isn't a moral difference-maker, or when sentience is uncertain, or estimates are low, or zero, but an agent might think that entity deserves moral consideration. Ought we mine the moon for minerals to make batteries? Inject aerosols into the stratosphere to reduce global warming? Dump radioactive waste into rivers? Raze parkland to build condominiums? Without minded entities being directly affected, sentientism has little to say about these actions, or must revert to instrumental

reasoning centered on the effects of actions on sentient entities. This means that the sentientist must always respond to moral dilemmas using a limited and sometimes unfitting logic. That logic is wider than anthropocentrism, but it cannot extend to some non-trivial morally charged choice situations to which moral philosophy should be responsive.

4. *Moral Worlds*

I hope to have motivated the idea that although sentientism is helpful for reasoning about the moral circle, it is unhelpful in cases involving minimally or non-minded entities. That doesn't make it a bad theory, but it is limited. Sentientism delivers strong moral verdicts in cases where empirical investigation or subjective estimation suggests that an entity has the capacity for suffering, but is silent in others where the morality of an action hinges on different morally relevant features of situations or entities. Thus, I contend sentience is sufficient, but not necessary for moral status.

Sentience is one of many *hinge properties* that can organize moral reasoning. Hinge properties function as deep-seated, effectively undeniable presuppositions in a reasoning process. The idea of a hinge property is inspired by Wittgenstein (1969), who thinks that some propositions are exempt from doubt within a practice:

> "…the *questions* that we raise and our *doubts*, depend on the fact that some propositions are exempt from doubt, are as it were like hinges on which those turn." [*sic.*] (§341, 44e).

Leading up to this characterization, Wittgenstein discusses the proposition "I know that I have two hands" as an example of a hinge proposition that serves as a foundational belief because "I am completely convinced of it" (§245-246), and that doubting such as proposition would be incoherent within my system of thought (§245-246). Recent work in *hinge theory*, which grows out of this notion, have addressed how hinges play a role in deep moral disagreement and other cognitive and practical phenomena (Caprioglio Panizza 2025). Hinge theory explains intractable moral disagreement as an exclusive disjunction of agents' foundational beliefs and values.

It is hard to imagine changing a hinge belief about having hands, and many moral hinges are indeed deeply held convictions, but I think moral hinges are malleable, and *should* be, especially in situations that challenge our foundational beliefs but also require us to be rational. Some contexts can drive us to *change* our moral beliefs. Wittgenstein himself says:

> "But what men consider reasonable or unreasonable alters. At certain periods men find reasonable what at other periods they found unreasonable. And vice versa…One cannot make experiments if there are not some things that one does not doubt." (§336-337)

Moral reasoning structured around a hinge property becomes inert in decision situations that outrun the hinge. This destabilizes reasoning about action under a framework in those choice situations. In these situations, I propose an agent should identify different hinge properties and take up alternative moral frameworks to reason coherently, and can do so through a process I will call *focal reasoning*. Focal reasoning *centers* different hinge properties and takes up the alternate moral frameworks that follow when choice situations warrant it—when default moral frameworks fail.

Consider a model of agents committed to different moral frameworks and hinge properties contemplating the action DESTROYING REDWOODS— a sentientist, a biocentrist who takes biological life as a hinge, and a rationalist who takes the ability for advanced reasoning and language as a hinge. Using the above decision theorem, relativized to different hinges and assuming risk aversion, we see how subjective estimates of hinge properties under different moral frameworks modulate expected utilities:

DESTROYING REDWOODS

Agent	**Moral framework**	**Hinge estimate**	**Utility if hinge present**	**Utility if hinge absent**	**Hinge-adjusted EU**	**Risk attitude**	**Attitude-adjusted EU**
1	Sentientism	0	-20	10	10	Averse	10
2	Biocentrism	1	-20	10	-20	Averse	-20
3	Rationalism	0	-20	10	10	Averse	10

Only the biocentrist can stand against DESTROYING REDWOODS, while the sentientist and rationalist must explain away the positive utilities their frameworks assign to the action, or permit it. The latter seems unpalatable from a naturalist point of view[3], and we can imagine structurally cases in which different moral frameworks will fall short as decision aids.

Sentientism may falter in some decision contexts, but does provide an immediate sorting of the world's entities into sentient and non-sentient categories, admitting the sentient category into the moral circle. Sentientism, however, is only one way to draw the moral circle. As a *conceptual edge* meant to carve out classes of entities in the world that are owed moral consideration, we can in principle carve up *moral space* differently.

If we imagine moral space as a set of possible worlds containing moral frameworks, entities and actions[4], we can think about ethical reasoning as a parsing those worlds—understood as possible counterfactual situations (Stalnaker 2003, Kripke 1980)—as morally acceptable or unacceptable. At each possible world, we can hold different elements constant and vary the others to examine different combinations of situations, frameworks, actions, and entities to explore moral space. At this juncture, one could also use decision theory to make probability-weighted comparisons among

[3] This intuition is admittedly environmentalist, but one experience in Muir Woods would make anyone think twice about razing it. See Muir 1908 for a flavor of naturalist attitudes.

[4] In the parlance of epistemology, this would be a function from actions-toward-entities and frameworks to permissibility—a computation that takes in actions-toward-entities and returns a permissibility verdict.

worlds to determine what we ought to do. I want to offer a coarser-grained picture on which every choice is a selection problem in which an agent chooses to realize some situation for an entity over other possibilities based on their beliefs and attitudes, and some of those possibilities might be better or worse for the entity. Moral constraints from different ethical frameworks may open and close off possibilities to the agent depending on their beliefs and attitudes about the moral standing of an entity, and the strength of their commitments to an ethical framework. Further, those beliefs and commitments will deliver different verdicts on how an agent ought to act with respect to different entities.

One way to think about the beliefs of a moral agent is through the lenses of different theories of moral standing, e.g. sentientism, biocentrism, rationalism, or ecocentrism (see Sebo 2026 for an overview of moral frameworks). These different moral stances will frame choice situations differently for that agent. The aim of *focal reasoning* is to make those alternate framings salient in moral deliberation and decision making, and in philosophical analysis. Depending on the way an agent takes the moral world to be, that is, which moral framework or hinge property they take up, that theory will shape their moral reasoning and result in different resolutions to decision problems across choice contexts.

Choosing *one* theory of moral status isn't something moral agents *must* do, the moral frameworks needed to resolve decision problems may vary depending on the

moral scene—the entities involved and the actions being considered. I'll sketch out this pragmatic ethical picture further using possible worlds and a two-dimensional style model of moral rationality. The idea is to lay out the relations between entities and what I'll call *moral worlds,* or contexts in which different moral frameworks are presupposed, to show how different moral presuppositions yield different permissibility verdicts across choice situations for different actions, and further, that choosing an ethical framework fitting for a decision situation is an important parameter and degree of freedom in decision making.

Consider a model of moral space that represents entities *E1…E4*, the generic action HARMING, and moral theories *T1…T4*, that rule on permissibility (P) and impermissibility (I) of an action ϕ in a world under a theory. We might call this model an *action space* that two-dimensionalizes entities on the vertical and moral standing relative to different hinge properties on the horizontal to visualize the set of moral worlds at which actions are morally evaluated:

ϕ = HARMING

		Sentience	***Rationality***	***Biology***	***Ecology***
		T1	***T2***	***T3***	***T4***
Human	***E1***	I	I	I	I
Octopus	***E2***	I	--	I	I
Redwood	***E3***	P	P	I	I
AI System	***E4***	--	--	P	--

In this action space, sentientism rules that harm is impermissible if an entity is sentient, and permissible if it is not, but cannot rule (--) on cases where sentience is absent or disputed. If an agent believes sentientism, it puts negative *moral pressure* on harming sentient entities but permits it toward non-sentient entities. Varying moral frameworks by looking across rows—considering different scopings of the moral circle—the permissibility of HARMING changes while the entity stays constant. Rulings on different entities under the same moral theory are visible by moving down the columns. For instance, moral worlds show that it is impermissible for the biologist or ecologist to harm a redwood, but permissible for the sentientist and the rationalist. If harming an entity is impermissible at all moral worlds, it's in the moral circle. If an action yields an impermissible space of moral worlds, the action is overdeterminately impermissible. The AI case is interesting, since there looks to be no open path for moral circle inclusion on present assumptions. That might be true, a different moral theory might be needed, or an existing moral theory's presuppositions might need adjusted to account for this choice situation. Another feature of focal reasoning and moral worlds is to make these kinds of ethical challenges salient.

Taking the choice situation about HARMING from the above action space, the focal reasoner admits that to rule on whether harming an entity is permissible, one must make a moral framework commitment *for that decision context*. If an agent wants to assert that an entity has moral status, they can justify it by centering a framework that

enables that assertion. Making moral space explicit helps bring out the downstream implications of a particular moral stance, which can aid in reasoning about how effective a particular moral theory might be against others in dealing with different possible cases. If there are surprising or unpalatable results in action space modeling, then we may need to reconsider whether we want to commit ourselves to a single theory of moral standing, or be open to the possibility that different theories may do better moral work in different contexts. Focal reasoning embraces this fact by switching moral worlds when needed.

With moral worlds, we can see that the generic action HARMING, yields more impermissible worlds than permissible worlds. From a focal reasoning perspective, this proliferation of impermissible verdicts in a plurality or majority of decision contexts puts negative moral pressure on the choosing to perform the action. The more resonance among moral theories about permissibility, the more confidence an agent can have in the morality of their action. Choosing to act against the array of moral worlds of an action space is an open option for an agent, but such a decision would be morally criticizable. At the case level, focal reasoning also allows for shifting one's thinking about an action under different *moral lenses*. An agent can change their moral lens and consider different framework-based verdicts against one another to decide how to act. Agents who consider the myriad of ways their action might be morally evaluated can get a better grip on its general permissibility, thus enriching their moral reasoning and

breaking them free of the limitations of any single ethical framework. Thus, moral worlds provide a systematic method for moral reasoning and analysis that avoids the drawbacks of steadfast commitment to a single theory or moral principle.

Under a single moral theory, a limited range of permissible actions are open to an agent. However, there might be permissible actions by the lights of different moral frameworks that fit with the sensibilities of a moral agent in a choice situation. A sentientist, for instance, may desire to grant some moral consideration to plants, ecosystems, or non-minded entities, but their framework is inflexible. Focal reasoning says to visit other moral worlds when faced with such inflexibility, and reveals alternate paths through moral space. This kind of pluralism about moral reasoning can generate different constraints and permissibility verdicts for an agent can consider in deciding to act, and so calibrates decision making to choice situations.

Moral worlds and focal reasoning highlight how different moral stances will rule on permissibility, and how those verdicts will vary depending on the choice context. Whether or not we can *know* which way the moral world *is*—that is, know which theory of moral grounding is *correct*—I want to stay neutral on, although I am sympathetic to the idea that there might be no fact of the matter. Wittgenstein (1965) held a view suggestive of this attitude toward moral truth:

> "Ethics so far as it springs from the desire to say something about the ultimate meaning of life, the absolute good, the absolute valuable, can be no science. What it says does not add to our *knowledge* in any sense. But it is a document of a tendency in the human mind which I personally cannot help respecting deeply and I would not for my life ridicule it." (11, my italics)

In any case, by mapping things out the two-dimensional way I've sketched, we can get a sense for the merits and sacrifices of different moral stances, see how different theories carve up moral space, and consider what alternative frameworks say we ought to do. As moral agents, we are then enabled to choose the framework that works best for a decision in context. One can have a default moral framework, but the spirit of focal reasoning is to make salient alternate paths for moral deliberation when that framework is uninstructive or inflexible.

5. *Objections*

Decision theory: This approach looks like it would deliver the same results as a decision theory that compares expected utilities associated with different moral frameworks. The moral worlds and focal reasoning approach differs from decision theory-based reasoning because an agent navigating moral worlds is deliberating at the level of their attitudes and ethical commitments, rather than decision theoretic expected

utility calculations. These pictures may be duals, or perhaps even interdefinable, but are nevertheless different modes of thinking about action and morality. Using decision theory to pick a utility winner is a fine-grained, formal epistemological way to reason about what to do using credence and expected utility. Moral worlds is a coarse-grained, more traditional epistemological approach operating at the level of beliefs, desires, and other attitudes. Both can account for action and decision, but they are different pictures.

Context sensitivity: Moral worlds invite the idea that ethics is context sensitive and that there aren't any moral facts of the matter because each choice situation can be interpreted and evaluated in different ways. It's not obvious to me that there *must* be deep moral truths or facts of the matter, when what matters morally is the *justification* of the performance of an action and its soundness given the conditions in which the action is performed and its impact on the world. A context sensitive dimension of ethical reason doesn't imply that ethical claims can't be true or false—it only implies that whether they come out true or false is a function of the choice situation, and a major parameter is a moral agent's presupposed ethical framework.

Relativism: If an agent can take up any moral framework in a decision situation, morality looks relative, and if no particular moral theory is correct, ethics collapses into a field of incommensurable and incomparable frameworks that agents can choose from depending on their beliefs and preferences. Moral worlds and focal reasoning promote principled *flexibility,* not an anything-goes relativism, because it seeks to provide

alternate strategies for moral reasoning when an agent's default framework falls short. A focal reasoner should switch to alternate moral frameworks that are *relevant* to a choice situation, and they may even have a preferred subjective ordering or frameworks, e.g. rationalism → sentientism → biologism, moving along frameworks when they fail to aid decision. An agent's intentions will matter here—if I intend to switch to rationalism to justify the mass slaughtering of animals, this kind of focal reasoning is itself subject to moral scrutiny.

Incoherence: If a moral agent were to adopt different moral attitudes across different choice situations, they would have inconsistent beliefs and commitments across contexts, making them structurally irrational. It's not clear to me that having different beliefs in different contexts means that an agent is structurally irrational. Their moral attitudes can be *context relative.* Philosophers of language discusses a structurally similar situation in relation to Frege puzzles: Lois Lane believes that Superman can fly, but that Clark Kent cannot. Superman is Clark Kent, suggesting that Lois is incoherent. Some replies are that Lois knows the same individual under different *guises* (Williamson 2024), or that her attitude can be interpreted non-uniformly over contexts (Dorr 2014). When Lois sees a man soaring above the city, she knows him as Superman, and when she sees a bespectacled and working on news articles, she knows him as Clark Kent—without being structurally irrational. I want to suggest a structural similarity in the focal reasoner in different decision contexts. An agent can operate

under the guises of different moral frameworks *as they are demanded by the choice situation* to reason pragmatically when their trusted moral concepts are unhelpful. As we can interpret a knower as rational when they track individuals under different guises, we can interpret a moral agent as rational when they consider moral permissibility under different moral frameworks in different decision contexts.

6. Conclusion

I have argued that sentientism is ill-equipped to guide moral reasoning and decision making in some non-trivial morally charged contexts, especially those with high uncertainty about sentience or low sentience estimates, through decision theoretic examinations. In response, I've developed moral worlds and focal reasoning as a coarse-grained philosophical approach to moral decision making and cognitive modeling that allows an agent to flexibly take up different moral frameworks to deliberate about the permissibility of actions in different choice situations. The upshot is that moral agents are not necessarily bound to a single moral theory or moral principle in reasoning about action, but can choose from a plurality of moral theories to reason about what they ought to do in different decision contexts.

References

Anthropic—Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey (2026). Emotion concepts and their function in a large language model. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/emotions/index.html.

Andrews, Kristen, Jonathan Birch, and Jeff Sebo (2025) Evaluating Animal Consciousness. *Science* 387 (6736): 822-824.

Barry, Christian and Garret Cullity (2022). Offsetting and Risk Imposition. *Ethics* 132 (3): 352-381.

Bentham, Jeremy (1907) *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press.

Birch, Jonathan (2024). The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. Oxford: Oxford University Press.

Birch, Jonathan, Donald M. Broom, Heather Browning, Andrew Crump, Simona Ginsburg, Marta Halina, David Harrison, Eva Jablonka, Andrew Y. Yee, François Kammerer, Colin Klein, Victor Lamme, Mattias Michel, Françoise Wemsfelder, and Oryan Zacks (2022). How should we study animal consciousness scientifically? *Journal of Consciousness Studies, 29*(3-4), 8–28.

Birch, Jonathan (2017) Animal sentience and the precautionary principle. *Animal Sentience* 16(1).

Buchak, Lara (2019). Weighing the Risks of Climate Change. *The Monist* 102: 66-83.

Butlin, Patrick, Robert Long, Tim Bayne, Yoshua Benigo, Jonathan Birch, David Chalmers, Axel Constant, George Deane, Eric Elmozino, Sephen M. Fleming, Xu Ji, Royota Kanai, Colin Klein, Grace Lindsay, Mattias Michel, Liad Mudrik, Megan A.K. Peters, Erix Schwitzgebel, Jonathan Simon, and Rufin VanRulen (2025). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences* Nov 10: S1364-6613 (25) 00286-4.

Caprioglio Panizza, Silvia. (2025). Deep Moral Disagreement and Unthinkable Possibilities. *International Journal of Philosophical Studies*, 1–20.

Chalmers, David J. (forthcoming). Sentience and Moral Status. In *The Importance of Being Conscious,* ed. Geoffrey Lee & Adam Pautz. Oxford University Press.

—— (2022) *Reality+: Virtual worlds and the philosophy of mind*. New York: W. W. Norton & Company.

Clatterbuck, Haley and Bob Fischer (2025). Navigating Uncertainty About Sentience. *Ethics* 135: 2.

de Araujo, Ivan E., Albino J. Oliveira-Maia, Tatyana D. Sotnikova, Raul R. Gainetdinov, Marc G. Caron, Miguel A.L. Nicolelis, and Sidney A. Simon (2008). Food Reward in the Absence of Taste Receptor Signaling. *Neuron* 57: 930-941

Muir, John (1908). The Hetch Hetchy Valley. *Sierra Club Bulletin* 6 (4).

DeGrazia, David (1998). Suffering. In *The Routledge Encyclopedia of Philosophy.* Taylor and Francis.

Dorr, Cian (2014). Transparency and the Context-Sensitivity of Attitude Reports. In.) *Empty Representations: Reference and Non-Existence* Manuel García-Carpintero & Genoveva Martí (eds.). Oxford: Oxford University Press.

Fischer, Bob, Joe Gottleib, Alexandra K. Schnell, and Meghan Barrett (2025). Defending and refining the Birch et al. (2021) precautionary framework for animal sentience. *Animal Welfare* 34: e28.

Frohlich, Joel and Tim Bayne (2025). Markers of consciousness in infants: Towards a 'cluster-based' approach. *Acta Paediatrica,* 114 (2): 285–291.

Kripke, Saul A. (1980) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

LeDoux, Joseph, Jonathan Birch, Kristen Andrews, Nicola S. Clayton, Nathaniel D. Daw, Chris Frith, Hakwan Lau, Megan A.K. Peters, Susan Schneider, Anil Seth, Thomas Suddendorf, and Marie M.P. Vandekerckhove (2023). Consciousness beyond the human case. *Current Biology* 33, 16: 832-840.

Lewis, David (1976). Probabilities of Conditionals and Conditional Probabilities. *The Philosophical Review* 83 (3): 297-315.

Nagel, Thomas (1974). What is it like to be a bat? *Philosophical Review* 83(4): 435–450.

Passos-Ferreira, Cláudia (2024). Can we detect consciousness in newborn infants? *Neuron* 112 (10):1520-1523.

Pennartz, Cyriel M.A., Michele Farisco, and Kathinka Evers (2019) Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach. *Frontiers in Systems Neuroscience* 13: 25.

Sebo, Jeff (2026). Moral Circle Explosion. In *The Oxford Handbook of Normative Ethics*, D. Copp, C. Rosati, and T. Rulli (eds.).

—— (2025). *The Moral Circle: Who Matters, What Matters, and Why*. New York: W. W. Norton & Company.

—— (2023). The Rebugnant Conclusion: Utilitarianism, Insects, Microbes, and AI Systems. *Ethics, Policy & Environment* 26 (2): 249-264.

—— (2018). The moral problem of other minds. *The Harvard Review of Philosophy* 25: 51–70.

Stalnaker, Robert C. (2003) Possible Worlds (1976/1984). In *Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays*. Oxford: Oxford University Press.

—— (1972). Letter to David Lewis. In *Ifs,* W.L Harper, R. Stalnaker, and G. Pearce (eds.).

Williamson, Timothy (2024) *Overfitting and Heuristics in Philosophy*. New York: Oxford University Press.

Wittgenstein, Ludwig (1969). *On Certainty.* New York: Basil Blackwell.

—— (1965). I: A Lecture on Ethics. *The Philosophical Review*, 74 (1): 3–12.