The Power of Rational Reflection

Javon Hickmon

June 2025

Contents

1 Introduction 1

2 Street's Reflective Equilibrium 2

3 Rawls' Reflective Equilibrium 3

4 The Limits of Rational Reflection 5

5 An Answer to the Possibility of Influence 10

6 An Answer to the Independence of Future Judgments 11

7 What about the Darwinian Dilemma? 12

1 Introduction

Within the debate surrounding Evolutionary Debunking Arguments (EDAs), the argument proposed by Sharon Street within *A Darwinian Dilemma for Realist Theories of Value* emerges as a major work aiming to debunk the theory of moral realism through an evolutionary approach. Within such arguments, *reflective equilibrium* is a commonly used term typically used to describe the state of a person's beliefs after rational reflection. A major issue, however, is that this term is often invoked without a consistent definition.

John Rawls, the philosopher who first coined the term, separated reflective equilibrium into two forms, *wide* and *narrow* reflective equilibrium. Each has distinct limitations that complicate their use in philosophical arguments, particularly in Street's EDA. I argue that Street's characterization of reflective equilibrium is invalid: she claims that our initial evaluative judgments constrain the outcome of reflective equilibrium, thereby limiting the scope of rational moral reflection.

This paper argues that such a constraint is unjustified. A truly rational reflective agent must be able to arrive at evaluative judgments that are not merely shaped or contaminated by initial intuitions, tendencies, or proto-beliefs. To support this claim, I present a method by which a rational agent, operating within Street's own framework, can transcend the limits she proposes and arrive at judgments that are genuinely independent of initial biases.

2 Street's Reflective Equilibrium

The core of Street's argument focuses on debunking the realist theory of value. The realist view of value is that "there are at least some evaluative facts or truths that hold independently of all our evaluative attitudes" (Street 2006, 110), where evaluative attitudes are later defined to include desires, attitudes of approval and disapproval, unreflective evaluative tendencies, and conscious/unconsciously held evaluative judgments. This definition is central to her argument because the EDA explores how our evaluative attitudes influence our evaluative judgments, especially with regard to the set of possible judgments upon rational reflection. Specifically, "Where anti-realists part ways with realists is in denying that there are evaluative truths which hold independently of the whole set of evaluative judgments we make or might make upon reflection, or independently of the whole set of other evaluative attitudes we hold or might hold upon reflection" (111).

The anti-realist view focuses on reflection as a means by which an agent can explore the space of possible evaluative judgments, moving past their initial unreflective evaluative tendencies. My argument does not aim to further explore or deny the anti-realist position; rather, I aim to examine and modify how Street has defined reflective equilibrium to understand how this augments her argument.

Street's definition of reflective equilibrium is clearly presented within the following passage:

> *The widespread consensus that the method of reflective equilibrium, broadly understood, is our sole means of proceeding in ethics is an acknowledgment of this fact: ultimately, we can test our evaluative judgements only by testing their consistency with our other evaluative judgements, combined of course with judgements about the (non-evaluative) facts. Thus, if the fund of evaluative judgements with which human reflection began was thoroughly contaminated with illegitimate influence - and the objector has offered no reason to doubt this part of the argument - then the tools of rational reflection were equally contaminated, for the latter are always just a subset of the former. It follows that all our reflection over the ages has really just been a process of assessing evaluative judgements that are mostly off the mark in terms of others that are mostly off the mark... So long as we assume that there*

> *is no relation between evolutionary influences and evaluative truth, the appeal to rational reflection offers no escape from the conclusion that, in the absence of an incredible coincidence, most of our evaluative judgements are likely to be false.* (Street 2006, 124–125).

The goal behind this line of reasoning is that she aims to show how our unreflective evaluative tendencies influence all our evaluative attitudes, even those within reflective equilibrium. Succinctly put, “My point here is instead the simple and plausible one that had the general content of our basic evaluative tendencies been very different, then the general content of our full-fledged evaluative judgements would also have been very different. . . ” (120). This builds to the core of her EDA, the two-horned 'Darwinian dilemma' for value realism.

Street writes that if our evaluative attitudes were to track evaluative truth, then one of two things must have happened:

1. It occurred by pure coincidence.
2. By some 'tracking account, ' evolutionary forces allowed our evaluative attitudes to track evaluative truth.

I focus on the second. The objection she brings to this horn asserts that by the 'adaptive link account,' our evaluative tendencies must be entirely explained by evolutionary factors. This account is combined with her definition of reflective equilibrium to form her refutation of the horn. She describes that the space of attainable evaluative attitudes for a rational reflective agent is limited by the adaptive link account.

It is important to note that Street does understand that our rational capabilities allow us to step back from a problem and consider it from different angles. ". . . I do not mean that we automatically or inevitably accept the full-fledged evaluative judgements that line up in content with our basic evaluative tendencies. Certainly not. For one thing, other causal influences can shape our evaluative judgements in ways that make them stray, perhaps quite far, from alignment with our more basic evaluative tendencies. For another thing, we are reflective creatures, and as such are capable of noticing any given evaluative tendency in ourselves, stepping back from it, and deciding on reflection to disavow it and fight against it rather than to endorse the content suggested by it” (120).

A major issue is that Street’s definition of reflective equilibrium is not consistent with John Rawls’ original definition of the term, as I will continue to explore.

3 Rawls’ Reflective Equilibrium

John Rawls first coined the term 'reflective equilibrium' within the paper *Justice as Fairness.* He describes it as the following: “By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium" (Rawls 1958, 20). By this definition, Rawls specifically clarifies that this form of reflective equilibrium simply aims to achieve internal coherence among our principles and judgments; however, he adds that it may not necessarily be stable.

This definition of reflective equilibrium thus far does not conflict with Street's main line of argument, since she primarily aims to show how our unreflective evaluative tendencies influence our later evaluative judgements. The contention comes when Rawls expounds upon this definition. He explores the implications embedded within the instability of reflective equilibrium. For a rational reflective agent, the equilibrium is liable to be upset by further examination or additional information, which could lead to a revision of judgment. With this instability, the core idea behind rational reflection is to justify our own convictions.

"When a person is presented with an intuitively appealing account of his sense of justice (one, say, which embodies various reasonable and natural presumptions), he may well revise his judgments to conform to its principles even though the theory does not fit his existing judgments exactly. He is especially likely to do this if he can find an explanation for the deviations which undermines his confidence in his original judgments and if the conception presented yields a judgment which he finds he can now accept" (48).

Rawls states that through rational reflection, a moral agent may augment their unreflective evaluative tendencies to favor another judgment. Here, he describes an intuitively appealing judgment, but it doesn't necessarily need to be so. He puts it more clearly in his later work, *The Independence of Moral Theory*: "Although in order to get started various judgments are viewed as firm enough to be taken provisionally as fixed points, there are no judgments on any level of generality that are in principle immune to revision" (Rawls 1974, 8). Under the Rawlsian reflective equilibrium, every judgement may be revised, even those that start as being influenced by unreflective evaluative tendencies. This is an extremely important distinction because Street's reflective equilibrium bakes in the assumption that the space of possible evaluative attitudes for a rational reflection agent cannot be synonymous with the space of all possible evaluative attitudes due to the limits of the agent's initial unreflective evaluative tendencies.

Despite these characteristics, Rawls goes on to acknowledge that there are multiple interpretations of reflective equilibrium. "For the notion varies depending upon whether one is to be presented with only those descriptions which more or less match one's existing judgments except for minor discrepancies, or whether one is to be presented with all possible descriptions to which one might plausibly conform one's judgments together with all relevant philosophical arguments for them" (Rawls 1958, 49). These two notions are what he later separates into *wide* and *narrow reflective equilibrium.* Narrow reflective equilibrium is when one just aims to reach internal coherence among their beliefs, while wide reflective equilibrium is where the agent brings in many other considerations (specifically those that are very different from one's existing judgments), ideally, all possible considerations.

Street's definition was much closer to narrow reflective equilibrium; however, Rawls expressly did not view narrow reflective equilibrium as enough to do more than smooth out bumps in the internal coherence of a belief. Despite this disagreement, even Rawls was skeptical regarding the practicality of wide reflective equilibrium: "To be sure, it is doubtful whether one can ever reach this state. For even if the idea of all possible descriptions and of all philosophically relevant arguments is well-defined (which is questionable), we cannot examine each of them" (49). In this statement, Rawls agrees with Street, where "The most we can do is to study the conceptions of justice known to us through the tradition of moral philosophy and any further ones that occur to us, and then to consider these" (49). Essentially, we are limited by that which we have/can comprehend. I argue this is not necessarily true, and for certain subclasses of problems, a rational reflective agent can feasibly consider all possible conceptions within the space of potential evaluative attitudes, resulting in an evaluative judgement that is not impacted by initial unreflective evaluative tendencies. I will in turn show how this impacts Street's EDA.

4 The Limits of Rational Reflection

To understand how I refute Street's formulation of reflective equilibrium, we represent her definition as a Markov process, where each decision made by agent A is an event within the continuous-time Markov chain. We can effectively utilize Markov chains for this formulation because each step within reflective equilibrium could be considered a stochastic process. Before the reflective process begins, a rational reflective agent **must** have some probabilistic likelihood of selecting evaluative judgements within the space of all possible judgements, even if the likelihood is zero. Markov chains focus on the likelihood of a future event occurring based on the current state, and this is exactly the process described by both Rawls and Street. In addition, the Markov chain is continuous-time because the process of reflective equilibrium is hardly discrete, meaning that there are not always set start and end times to a process of rational reflection.

To effectively consider Street's argument, I break it down into its formal logic:

- Let A be a rational reflective moral agent.

- Let $p \in \mathcal{P}$ be the current value assignment proposition sampled from the set of all possible value assignment propositions $\mathcal{P}$.
- Let $t \in [0, \infty)$ be the current time-step
- Let α_p be the space of all possible evaluative attitudes for the given proposition p. Note both α and p are considered to be within the same space, but α may be aligned or misaligned with p.
- Let $Y_t \in \alpha_p$ be the evaluative attitude for agent A held at time t, where Y_t must be within α_p.
- Let $X_t \in Y_t$ be the evaluative judgment made at time t, where X_t must be within agent A's evaluative attitudes at time t.
- Let $(X_t)_{t \geq 0}$ be a continuous-time Markov chain of evaluations made within the process of rational reflection over proposition p.

Note for these definitions, I reference Street's original explanation of the terms: "Evaluative attitudes I understand to include states such as desires, attitudes of approval and disapproval, unreflective evaluative tendencies such as the tendency to experience X as counting in favor of or demanding Y, and consciously or unconsciously held evaluative judgements, such as judgements about what is a reason for what, about what one should or ought to do, about what is good, valuable, or worthwhile, about what is morally right or wrong, and so on" (Street 2006, 110). It is important to make the distinction that α is simply the space of all possible evaluative attitudes for the problem, not for agent A. This means that there may be evaluative attitudes contained within α_p that are impossible for A to consider.

Given the above, Street makes the following assertion:

$$\forall p \forall A \forall t \geq 0, X_t \perp X_0 \implies \mathbb{E}[X_t] = 0$$

In other words, Street states that any evaluations in contradiction to the initial unreflective evaluative tendency X_0 cannot be selected due to the impact evolutionary forces have on our tendencies. For this, I once again reference Street's writing, "Thus, if the fund of evaluative judgements with which human reflection began was thoroughly contaminated with illegitimate influence... then the tools of rational reflection were equally contaminated..." (124).

For the sake of argument, I focus on the second statement, "then the tools of rational reflection were equally contaminated." With Street's formulation thoroughly described, I begin my refutation by clarifying what it means for there to exist a rational reflective agent A. For a value-assignment proposition p that has not been considered (so we start at time $t = 0$):

1. A rational reflective agent A must be equally likely to select given evaluative judgments X or Y if A evaluates them to have equal value. For

example, a student would be equally likely to judge test X or test Y as easy if they were confident in all the material from the course.

2. A fair rational reflective moral agent A must have some non-zero likelihood of selecting evaluative judgement X if it has not been evaluated, or if given an adequate reason. Furthermore, A must be more likely to select judgment X rather than Y if they deem the rationale for X to be stronger than that for Y. This clause simply clarifies that a fair agent should be open to making an informed decision before starting the rational reflection process.

3. A rational reflective agent A must have a zero likelihood of selecting an already considered and rejected evaluative judgement X without the presence of new evaluative information.

4. A rational reflective agent A must assign equal value to evaluative judgements that are not relevant to A

To refute Street's original statement, I only need to show that there exists an agent A that can rationally come to a judgment that is in direct contradiction to their original evaluative tendency. In other words:

$$\exists p \land \exists A, \exists t \geq 0 \text{ such that } (X_t \perp X_0) \land (\mathbb{E}[X_t] \neq 0)$$

My argument roughly follows this structure:

- Premise 1: A rational reflective agent can make an evaluative judgment that directly contradicts their initial unreflective evaluative tendencies.
- Premise 2: A rational reflective agent can explore the full space of evaluative attitudes α (i.e., wide reflective equilibrium).
- Conclusion: Therefore, a rational reflective agent can arrive at an evaluative judgment in wide reflective equilibrium that contradicts their initial unreflective evaluative tendencies.

Considering a value assignment proposition p, let there exist some unbiased rational reflective moral agent A.

To address premise 1, I focus on clause 2 of the definition of rational reflective moral agent. Specifically, if A is provided a reason that they deem to be strong enough, then they can be convinced that a different judgment is correct. For example, let's say the proposition p is "eating 100 pounds of ice cream is good for you." We are given that agent A is a child who, due to evolutionary pressures, ended up with taste buds that make it more likely for them to evaluate foods with high sugar content as tasting good. The child's initial unreflective evaluative tendency is that sweet things taste good, and whatever tastes good must be good for them; however, in the future, they would likely learn about the health issues that could be caused by eating so much ice cream so quickly. Thus, they would change their evaluative judgment without changing their tendency to think sweet things taste good.

Street is actually fully on board with this premise. "ultimately, we can test our evaluative judgements only by testing their consistency with our other evaluative judgements, combined of course with judgements about the (non-evaluative) facts" (Street 2006, 124). Non-evaluative facts are valid ways of examining our judgments, and thus, are one way to form a judgment that contradicts initial unreflective evaluative tendencies. The disagreement comes in premise 2.

To address premise 2, I focus on the space of all potential value assignment propositions $\mathcal{P}$. For any value assignment proposition p, there is an implicit relationship, where p only evaluates to true if $\forall q \in \mathcal{P}$ such that $q \neq p, p \wedge \neg q$. This is what I call the independence of proposition space. This distinction is important because within proposition space, there exist no subsets; rather, there are similarities. One can think of it like a mathematical unit vector space, where each proposition is a vector that could be compared to another vector by means of cosine similarity. This also means that a proposition like "kicking kittens is always wrong" is not a superset of the proposition "kicking kittens is sometimes wrong." They are independent propositions that are similar in their value assignment.

We can use another example to further understand the independence of proposition space. Given the proposition, "it is always good not to murder your children", p corresponds to endorsing the proposition that refraining from such an act is morally good. This implies the rejection of all alternative moral framings q that contradict, deny, or are indifferent toward that position because they are logically incompatible with p, and thus are excluded by $p \wedge \neg q$. An invalid q would be something like "it is morally neutral to murder your children." By the logical definition above, "it is always good not to murder your children" and "it is **not** morally neutral to murder your children" are logically consistent and so the statement "it is morally neutral to murder your children" would be invalid. One can think of this as a "rather than" relationship, where we are stating X is permissible rather than Y. Note that the independence of proposition space does not make any claims as to the truth or falsity of p, it solely maps p in terms of all other propositions within $\mathcal{P}$.

Finally, for premise 2, I must clarify what is meant by "evaluation" through "rational reflection". For rational reflection to occur, A must consider the reasons why p. One could argue that A must also consider all the reasons why p rather than q; however, due to the independence of p as defined above, such a consideration is not necessary. Therefore, evaluative judgement X_t is a valid evaluative judgement through rational reflection, if and only if it is formed by a proposition p and justified by some rationale r. Note that this rationale does not need to be strong, but *reason* must by definition be used within rational reflection. With this in mind, I update my set of definitions:

- Let $\mathcal{R}$ be the set of all **relevant** rationales for proposition p. Similarly $\mathcal{R}_n$ be the $\mathcal{R}$ at pass n when updating p.
- Let $r \in \mathcal{R}$ be a rationale from $\mathcal{R}$
- Let $X_t \in Y_t$ such that $X_t = r \vdash p$ be the evaluative judgment made at time t, where X_t must be within agent A's evaluative attitudes at time t, and such that X_t is the true evaluation of proposition p as justified by rationale r.

With this definition of "evaluative judgment by means of rational reflection", we start to see the challenges Rawls faced when examining wide reflective equilibrium. "To be sure, it is doubtful whether one can ever reach this state. For even if the idea of all possible descriptions and of all philosophically relevant arguments is well-defined (which is questionable), we cannot examine each of them" (Rawls 1958, 49). The full space $\mathcal{R}$ is intractable for a realistic agent A to evaluate, since it is theoretically countably infinite. Despite this, one aspect of the space that was unexplored by both Rawls and Street is the notion of relevance.

A rationale r is "relevant" to agent A **if and only if** X_t would be affected (either increased agreement or disagreement) as a result. This means that not all rationales are relevant to A. For instance, given the proposition "kicking kittens is wrong," the rationale "because I said so" would likely have a lower likelihood of being selected than "because it introduces unnecessary suffering into the world." This example creates numerous implications regarding how to quantify the likelihood of a rationale being selected; however, this is irrelevant to my argument. Rather, I solely aim to show that certain rationales have a zero likelihood of being selected because they are irrelevant.

Since the space of $\mathcal{R}$ is countably infinite, according to clause 4 of the definition of rational reflective agent, the likelihood of selecting each rational will approach zero. Mathematically:

$$\lim_{i \to \infty} \mathbb{E}[r_i] = 0$$

Therefore, if we are given a limited proposition p such that the majority of possible decisions are irrelevant, the space $\mathcal{R}$ is countably finite and agent A can feasibly engage in wide reflective equilibrium. In order to form this proposition, one simply needs to embed a rationale within p such that $p_2 = r \vdash p_1$. This works because the propositions and evaluative judgments are from the same space; a judgment simply affirms or denies a proposition. For example, given the proposition "kicking kittens is bad," the space of all possible rationales $\mathcal{R}_p$ is intractable, because {"because I said so", "because they are small", "because it introduces unnecessary suffering", ...} are all valid rationales, no matter their strength. Despite this, I can redefine the proposition to include one of the rationales, thereby limiting the space of relevant evaluative judgments that can be made.

An example of the new p would be ”kicking kittens is bad because it introduces unnecessary suffering.” We can effectively add this rationale to p because if A was already likely to agree with the statement ”kicking kittens is bad,” then ”because it introduces unnecessary suffering” would only increase their agreement with p so it is a relevant rationale. With this updated p the new space of relevant rationales is even more limited and would look like {”and because suffering is bad”, ”and because unnecessary suffering is gross”, ... }.

Note that for the second pass of relevant rationales, we may not use any rationale that was considered for the first pass. This is due to the fact that, by definition of relevant, the rationale would not augment A’s judgment toward the new proposition p.

These distinctions are important because recall that for my argument to work, I only need to show that there exists a proposition that satisfies the following criteria:

$$\exists p \wedge \exists A, \exists t \geq 0 \text{ such that } (X_t \perp X_0) \wedge (\mathbb{E}[X_t] \neq 0)$$

Therefore, I am able to redefine the proposition p to include the rationale r while still rejecting Street’s description of the limits of rational reflection. Thus, I have shown that there exists a rational reflective agent that can arrive at an evaluative judgment in wide reflective equilibrium, which contradicts their initial unreflective evaluative tendencies.

5 An Answer to the Possibility of Influence

A major objection to my line of argument ties back to Street’s original formulation. The objection goes as follows: for Street’s argument to get off the ground, she doesn’t need to show that one’s final state is *necessarily* influenced by our initial unreflective evaluative tendencies; she only needs to show that it *can* be influenced by these tendencies. The distinction is important because if she aimed to show that one’s final state is necessarily influenced by initial tendencies, then she must show this for all possible propositions p in the proposition space $\mathcal{P}$.

The primary issue is that this objection produces the skeptic’s conclusion to the Darwinian dilemma. If Street only aimed to show that one’s final state *can* be influenced by initial tendencies, then the conclusion would be to plant a seed of doubt that one’s judgments are completely unbiased from their initial tendencies. Despite this, Street’s argument is far more than a skeptic’s conclusion. As an EDA, the argument aims to debunk the realist theory of value by showing the necessity of influence. ”My conclusion: the content of human evaluative judgements has been tremendously influenced - indirectly influenced, in the way I have indicated, but nevertheless tremendously influenced - by the forces of natural selection, such that our system of evaluative judgements is saturated with evolutionary influence. The truth of some account very roughly along these lines is all that is required for the Darwinian Dilemma to get off the ground" (Street 2006, 121).

Therefore, my response to this objection is quite simple: Street's original goal was to show that one's final state is necessarily influenced by initial unreflective evaluative tendencies. Once again, "Thus, if the fund of evaluative judgements with which human reflection began was thoroughly contaminated with illegitimate influence... then the tools of rational reflection were equally contaminated" (124). This clearly shows that Street's focus was to show that all humans have had their tools of rational reflection contaminated; therefore, their future judgments must be influenced by the initial tendencies. She does this because her goal is to show that if any evaluative judgement can be influenced by initial tendencies, then we may lose confidence in the independence of all our evaluative judgements from these tendencies. This is how she forms the Darwinian dilemma. If one's final state can be influenced by initial judgments, then the only way for the realist to have confidence that their evaluative judgments are truth-tracking would be to have some tracking account; otherwise, natural selection would be considered purely a distorting force on our judgments.

6 An Answer to the Independence of Future Judgments

A second major objection to my argument relies on Street's original characterization of the influence of unreflective evaluative tendencies. Essentially, this contests that even if one were able to consider all possibilities within the proposition space through wide reflective equilibrium, their initial unreflective evaluative tendencies would influence any judgments they come to. This "corruption," as Street labels it, eliminates our ability to make a completely unbiased judgment.

This objection is important because it would mean that Street does not need to care whether the agent engages in wide or narrow reflective equilibrium. All of their future judgments would have been determined by initial tendencies. To take this line of reasoning to the extreme, one could conclude that any arguments about the process of reflection are meaningless, due to the overwhelming influence of initial tendencies.

To answer this objection, I turn back to how I defined a rational reflective agent, specifically clause 2 of the definition, "A **fair** rational reflective moral agent A must have some non-zero likelihood of selecting evaluative judgement X if it has not been evaluated, or if given an adequate reason." The extra term "fair" is only added here because neither Street nor Rawls explicitly requires that a moral agent; however, to counter this objection, I only need to show that an agent can escape this dilemma. If an agent is fair, then their rational reflection should fully consider relevant rationales.

If Street's argument was "had the general content of our basic evaluative tendencies been very different, then the general content of our full-fledged evaluative judgements would also have been very different…" (Street 2006, 120), then my counterargument is that for a fair rational reflective agent, the content of their basic evaluative tendencies do not ultimately decide their future judgments. Rational reflection has much more evaluative power than Street credits.

7 What about the Darwinian Dilemma?

Given the possibility of a rational reflective agent being able to enter wide reflective equilibrium, Street's Darwinian dilemma now must account for such a scenario. Evolution can no longer be considered to be a purely distorting force, or rather, it cannot be considered to be an irreparable distortion. The ability for a rational reflective agent to form evaluative judgments through which their initial unreflective evaluative tendencies have no effect shows that rational reflection is able to offset the influence of evolutionary pressures. I have not made any claims as to the truth-tracking ability of such reflection, which is outside the scope of this paper. Rather, for Street's argument to be strong, she must address how rational reflection can eliminate the impact of evolutionary pressures (as she describes them) for the class of propositions as described within this paper.

References

Rawls, John. 1958. "Justice as fairness." *The philosophical review* 67 (2): 164–194.

———. 1974. "The independence of moral theory." In *Proceedings and addresses of the American Philosophical Association,* 48:5–22. JSTOR.

Street, Sharon. 2006. "A Darwinian dilemma for realist theories of value." *Philosophical studies,* 109–166.