What is the Fallacy of Approximation?[*]

Matthew Hammerton and Sovan Patra

Suppose that you want to quit drinking. The experts advise that it is best to do *all* of the following:

(a) Start attending AA meetings.
(b) Stop hanging out with your alcoholic friends.
(c) Throw out (or, more prudently, sell) your whisky collection.

Unfortunately, the closest AA chapter is very far away and thus you cannot do (a). In these circumstances, you might reason as follows: even if I can't give myself the best chance of escaping alcoholism (since I can't do (a)), I can still give myself the next best chance by doing (b) and (c), since, by doing them, I would be in a situation that "most closely resembles" or "approximates" the ideal situation.

Now, consider a second scenario: suppose that at the end of a long work-day, you would ideally like to wind down with a cold beer. Yet, you forgot to transfer bottles from the carton to the fridge and thus, although there is enough beer, none of it is cold. In such a circumstance, you might reason that having room temperature beer is better than having no beer at all since it is a closer "approximation" to the ideal situation.

Finally, consider a third scenario (adapted from Sen 2009). Suppose you are throwing a party for friends, all of whom prefer red wine over white. While it would be ideal to serve all your guests red wine, unfortunately, there is not enough red wine for all. However, there is sufficient white wine. In such a situation, you might reason that the next best situation is one where everyone is served a mixture of red and white wine since a mixture of red and white is a closer "approximation"—in some sense, at least—to serving red, compared to serving only white wine.

At this point, it might be useful to compare our intuitions in these cases. We might judge that the person who chooses to dispose of his alcohol stash and

[*] This is the penultimate draft of an article that is published in *Erkenntnis*, and is available at https://doi.org/10.1007/s10670-022-00642-7. Please cite the published version.

ignore his alcoholic friends chooses rationally; we might be a little more circumspect about the rationality of choosing to drink warm beer over drinking something else; and, we might dismiss the decision to serve guests a mixture of red and white wine as absurd. Whatever we think of the rationality of the conclusions, we will surely observe that, in each of the different scenarios, the conclusion is justified by appealing to the claim that the chosen scenario approximates the ideal scenario better than any other alternative. Furthermore, whatever we think about the rationality of the conclusions, we might think that each of the conclusions has been poorly justified. More specifically, we might think that the alternatives that most resemble one another in their "features" are not necessarily, or even likely to be, closest in their desirability.

This idea—that *closeness in features* does not entail *closeness in desirability*—is often appealed to in moral and political philosophy.[2] Many call it "the problem of second best", but this wrongly suggests that it only applies to situations where one is interested in the second best option. To avoid this implication, we will follow Estlund (2019) and call it "the fallacy of approximation". This article develops a general account of the fallacy of approximation that improves on accounts currently available in the literature. In §1 we explore how many philosophical discussions of the fallacy of approximation appeal to the famous theorem of Lipsey and Lancaster. We argue that these appeals typically misuse the theorem and that a full account of the fallacy of approximation must be developed independently of Lipsey and Lancaster's theorem. In §2 we explore David Estlund's account of the fallacy of approximation—the *only* systematic account developed in the literature that moves beyond the Lipsey and Lancaster theorem. We argue that, although Estlund's account is an improvement on previous accounts of the fallacy in the literature, it has several serious flaws that make it untenable. Finally, in §3, we develop our own account of the fallacy of approximation which is based on the notion of "relevant descriptive similarity". By capturing the wide variety of contexts in which approximation reasoning is used, our account avoids the problems that undermine Estlund's account. It also shows that there is no simple

[2] For example, see: Elster (1986), Goodin (1995), Coram (1996), Räikkä (2000), Brennan and Pettit (2005), Pettit (2006), Brennan (2007), Margalit (2010), Cavallero (2011), Goodin (2012), Heath (2013), Räikkä (2014), Wiens (2016), Steinberg (2017), Berg (2019), and Estlund (2019).

"fallacy of approximation" but rather several different ways in which approximation reasoning can go wrong. Finally, it shows how reasoning by approximation is a legitimate form of reasoning in many circumstances. Given this, our account makes progress possible in philosophical debates that use approximation reasoning.

1. The Lipsey and Lancaster Theorem

Among philosophers discussing the approximation fallacy, there is a consensus that the first formalisation of the fallacy appears in the work of the economists Lipsey and Lancaster (1956). Lipsey and Lancaster present a theorem in mathematics (more specifically, optimisation theory) and apply that theorem to economics (more specifically, welfare economics), which can be interpreted in the following manner. Suppose the value of a variable, y, depends on the values of the real valued continuous variables, x and z, according to a function, $y = f(x, z)$. Suppose also that one is interested in finding the values of x and z that will maximise the value of y.

Further suppose that the solution to the optimisation problem is $x = x^*$ and $z = z^*$. In other words, suppose the starred values of the variables give the optimal value of y. Now, if $z = z^*$ is unrealisable (for whatever reason), the problem of finding the optimal value of y becomes a "constrained" optimisation problem, where one is looking for—as before—the values of x and z that define the optimal value of y but with the constraint that $z \neq z^*$. Suppose the solution to the constrained optimisation problem is $x = \hat{x}$ and $z = \hat{z}$. Lipsey and Lancaster's theorem can then be taken to establish, for any arbitrary function, f, firstly, that $\hat{z}$ *need not* be as close as possible to z^* and secondly, that $\hat{x}$ *need not* equal x^*.[3]

It is easy to see why many philosophers have taken this famous theorem as establishing a "fallacy of approximation". First, in some descriptive sense at

[3] To be strict, Lipsey and Lancaster show that, when a function is being optimised, given some constraints on the values of the variables being chosen to optimise the function, if at least one of the necessary conditions for an interior optimum cannot be satisfied, then this is a (these are) further constraint(s) on the original problem. They then prove that *even* the satisfiable first order conditions of the original problem *need not be identical to* the corresponding first order conditions of the new, "further constrained" problem. Wiens' (2016) account considers both Lipsey and Lancaster's formalisation (he calls it a "ratio" principle violation) and the formalisation presented above (he calls it a "level" principle violation).

least, $x = x^*$ and $z = \bar{z}$ (where $\bar{z}$ is the closest possible value to z^*) is the closest approximation to the optimal situation characterised by $x = x^*$ and $z = z^*$. However, from that descriptive resemblance, if one were to infer that $x = x^*$ and $z = \bar{z}$ determine the next best value of y given the unrealisability of $z = z^*$, the inference would be *invalid*. In other words, the next closest scenario to the optimal is not necessarily the next optimal. Second, the theorem, as presented here, justifies the two different ways in which the fallacy of approximation is presented in the philosophical literature. In suggesting that the next optimal situation is one where $x = x^*$, the "approximator" is appealing to the idea that the next optimal situation should satisfy *as many* of the conditions as there are in the optimal situation. Further, in suggesting that the next most optimal situation is defined by $z = \bar{z}$, the approximator is appealing to the idea that the next optimal alternative will satisfy the conditions realized in the optimal situation *to a greater degree* than any other alternative. That said, we will presently argue that the attempt to brand reasoning by approximation as fallacious, simply on account of the theorem of the second best is, itself, fallacious.

To do that we begin by identifying some key features of the discussion on the fallacy in the literature. Firstly, and curiously, before Estlund (2019) there was no attempt to give formal structure to reasoning by approximation independently of appealing to Lipsey and Lancaster's theorem. This lack of a standardised structure has had the undesirable consequence of spawning accounts of the fallacy that are vague with respect to *when* one can be considered to be reasoning by approximation; with respect to *why* reasoning by approximation is fallacious; and, with respect to *the conditions under which* reasoning by approximation might not be fallacious. Here are a few examples. Goodin (2012: 157) writes:

> [T]he Theory of Second-Best says this: If the first-best state of affairs cannot be obtained, the second-best state of affairs is not necessarily identical to the first-best in any respect. Whereas Lipsey and Lancaster's stronger version would say "necessarily not," the weaker and more general version that I shall be discussing says merely "not necessarily.

Here, firstly, Goodin misreads the mathematical theorem to establish that the second best state is necessarily not similar to the first best. The theorem does nothing of this sort: it does not provide a valid argument for a conclusion that the second best *cannot* resemble the first best in any way; instead, it shows that to argue that *it must* is invalid. So Goodin's "weaker and general" version is not any weaker or any more general than Lipsey and Lancaster's theorem. Secondly, it is also unclear what Goodin's putatively "more general" account is. The account is built on the example of "Goodin's car": if one's ideal car is a new silver Rolls, and if such a car is unavailable, one might have committed the approximation fallacy by preferring, say, a new silver Toyota over a not-so-old black Mercedes. He goes on to claim that one fallaciously reasons by approximation in this instance since one commits the "error of optimising on a subset of all the dimensions that are actually important to you" (2012: 159). In other words, the Toyota-chooser erred in deciding on the best available car by focussing on two criteria important to her (colour and newness) while neglecting the third (luxuriousness or "snob" value).

Two problems arise when we try to piece together Goodin's account from this example. Firstly, if luxury is important to the agent, it is unclear to us how Goodin can plausibly explain the Toyota chooser *forgetting* about luxury in concluding that the new silver Toyota has a closer resemblance to a new silver Rolls than the Mercedes. Secondly, and more crucially, Goodin's account is unclear about where the fallacy lies: is it in *misidentifying* something as the *next closest option to the ideal* or, is it in *wrongly inferring* that the (correctly identified) *next closest option* is the *next best option*?

Next, consider Brennan (2007: 124):

> The original formulation of the second-best theorem (Lipsey and Lancaster, 1956–7) relates to a circumstance in which the conceptual ideal is specified in terms of the simultaneous application of three interrelated conditions: the theorem states that if there is a constraint that prevents satisfying one of these conditions, then the feasible best (feasible given that constraint) will in general involve violating all three conditions.

Again, this representation of the theorem partially misreads it. The theorem does not claim that a constraint on one of the necessary conditions for (unconstrained) optimisation will *generally* lead to a *violation of all of the other necessary conditions* in the (constrained) optimisation problem: rather the theorem states that if one of the necessary conditions needed for optimality is violated, then *all* *of the other (unaffected) necessary conditions will not necessarily hold in the new optimal.*

These misreadings of the theorem and its scope have combined to produce, in the literature, a complacent dismissiveness of all forms of reasoning by approximation. The truth of the theorem has been incorrectly conflated with the fallaciousness of *all* appeals to approximation. It is important to note here that the theorem is a mathematical truth which does not make any reference to "approximation". So, unless one has a characterization of reasoning by approximation, one cannot appeal to this mathematical truth to render such reasoning fallacious. It is equally important to note that the theorem is true for a very specific set of (mathematical) assumptions. First, the function being optimised must be *differentiable*. Differentiability implies *continuity* (loosely, that, if the function is plotted, there will be no "breaks" in the graph)[4] and *smoothness* (loosely, that, if the function is plotted, the graph has no "pointy" parts). This immediately precludes its usage in examples such as Goodin's car—a function that has colour as an argument cannot plausibly be differentiable—as well as many other ordinary instances of reasoning by approximation. Second, the theorem does not apply if the objective and the constraint functions are linearly separable (loosely, a function is separable if the rate at which the dependent variable changes with respect to an independent variable, depends only on that independent variable).[5] Yet in some contexts where you might reason by approximation, the objective function is linearly separable, and in other contexts (such as the three examples we opened with) there is not enough information to determine whether it is linearly separable. Clearly then, it is a mistake to suppose that the theorem covers all cases of reasoning by approximation. At best, the theorem provides an *example* of a situation where it would be invalid to reason by approximation.[6]

[4] See, also, the discussion on pages 9-10.

[5] More technically, a function $y = f(x, z)$ is linearly separable iff $y = f(x, z) = g(x) + h(z)$. In the optimisation of linearly separable functions, if a necessary condition for optimisation is not met, then in the new optimum the other (unaffected) necessary conditions still hold. Subsequent literature has identified the separability conditions under which Lipsey and Lancaster's theorem no longer holds. For details, see Ng (2004: 195-196).

[6] Our observations attempt to correct the conflation of Lipsey and Lancaster's theorem with reasoning by approximation by showing that the *theorem only applies to particular types of objective/value functions*. Wiens (2020) identifies a related, but distinct, misunderstanding in

Despite this problem, the literature has at least done well to observe that (non-linear) interaction effects among value-contributing variables, or a discontinuous relation between values and value-creating variables, are instances where reasoning by approximation is fallacious. With reference to the first observation, in addition to Brennan above, here is Goodin (2012: 159):

> The second factor driving the Second-Best phenomenon is "interaction" across those dimensions. So, for example, education interacts with health which interacts with employment: the more education people have, the better able they are to make healthy lifestyle choices and the better able they are to take advantage of employment opportunities; and the healthier people are, the better able they are to hold down a job. That is why policymakers need to consider the entire suite of education-health-employment policies all together in a holistic manner, rather than just attending to them separately.

It is important to note why reasoning by approximation is fallacious when such "interaction" effects are neglected. Consider the following example adapted from Estlund (2019). Suppose that you are suffering from an ankle sprain and dry eyes. Your doctor prescribes an analgesic rub for the former and eye drops for the latter. Ideally, you would take both these medicines. However, if you are unable to take both because one of them is out of stock then at least taking one of the two (the next closest alternative to the ideal) would be best. Contrast this with a situation where you have been prescribed two pills—Pill A to suppress a critical heart condition and Pill B to mitigate the debilitating side effects of Pill A. Again, it would be ideal to take both. However, if one of these pills is unavailable, taking only the other one (the next closest alternative to the ideal) may not be such a good idea. The crucial difference in the two situations is that while, in the former, the remedies "independently" contribute to the desired objective (say, holistic

the literature about how broadly the theorem can be applied. The literature's overenthusiasm for applying the theorem, he says, arises from a failure to realise that the theorem is true *only if* a "specific type of *constraint* on realising the ideal" exists. So, while we agree with Wiens that there has, indeed, been a tendency to over-rely on the theorem, our reasons differ (but are not inconsistent). While we argue that the theorem is appealed to even when the *objective* function precludes such application, Wiens argues that the theorem is appealed to even though the violated *constraint* rules out such application.

health), in the latter, they "interact" to achieve that objective. It may be tempting to conclude from this example that all fallacious appeals to approximation involve some form of causal interaction between the value creating conditions. However, as Estlund recognizes, this would be an error. For example, consider again Goodin's car. A car's colour is causally unrelated to both its make and its luxuriousness. Yet, choosing the new silver Toyota *on the ground that* it most closely resembles a new silver Rolls would be fallacious reasoning. This example clearly illustrates that approximation-based reasoning can go wrong in a variety of ways (either because of false premises or because of a weak inference); as we show later, our account has the advantage of being able to diagnose, precisely, the flaw in a flawed instance of such reasoning.

The philosophical literature has also correctly recognised that a "discontinuous" relationship between the objective and the attributes contributing to the objective *can* render reasoning by approximation invalid. For example, Coram (2012: 94) draws this connection by diagnosing a related "fallacy of continuity", which he explains as follows:

> This is the idea that similar initial conditions will give similar results. In mathematics continuity roughly means that, if a function of x, say $f(x)$, is continuous, then two points close to each other are mapped close to each other. In other words, small changes in the xs do not lead to big jumps in the output, the $f(x)$s. In this case f can be thought of as the rules and x as the conditions. Thus a rules function f would map conditions into outcomes. It would be continuous if, by analogy, small changes in the initial conditions did not cause large changes in the output.

To see why "discontinuity" makes us more susceptible to the fallacy of approximation, suppose that you are considering making additional contributions to your retirement fund. Your financial advisor tells you that making an additional contribution is only advisable if the government will match it. Given current government policy, she therefore advises you to make a $1000 contribution. However, when the time comes to make this contribution you realise that you only have $900 available and decide to contribute this amount because it is pretty close to the ideal contribution of $1000. Whether this is a good decision will depend on the exact nature of the government policies that led to your advisor's recommendation. Consider two possible policies:

> CONTINUOUS: For any additional amount that a taxpayer contributes to their retirement fund, the government will match that amount, up to a maximum of $1000.
>
> DISCONTINUOUS: For any additional amount of $1000 or more that a taxpayer contributes to their retirement fund, the government will make a one-off contribution of $1000.

If CONTINUOUS is the government's policy, then contributing $900 was the best option in your circumstances, as the government will match your contribution. However, if DISCONTINUOUS is the government's policy then contributing $900 was a mistake, as the government will *not* match your contribution and unmatched contributions are financially imprudent. The moral of the story is that getting as close as possible to the optimal value of a variable may be desirable with a continuous relationship between the variable and the objective, but not with a discontinuous one.

2. Estlund's Account

In Chapter fourteen of his book *Utopophobia*, David Estlund develops the first systematic account of the fallacy of approximation. Unlike previous discussions of the fallacy, his account does not appeal to the theorem of Lipsey and Lancaster.[7] Instead, it is based on a principle he labels "Superset":

> For any valuable set of conditions S, and satisfaction of any subset of it ss (including the null set), any subset of S (including S itself) that is a proper superset of ss will be more valuable than satisfaction of ss. (2019: 274)

He then diagnoses the fallacy of approximation as follows:

[7] In this respect, Estlund's project is importantly different from that of Wiens (2020). Wiens shows that the *Theorem of Second Best* can be extended to "non-native" domains by modelling problems in these domains to mathematically resemble the context in which Lipsey and Lancaster proved the theorem. By contrast, Estlund is interested in a more general form of reasoning ("approximation reasoning") which is present even in contexts that don't mathematically resemble the context of Lipsey and Lancaster's theorem. When we offer our own account later, we will follow Estlund in having this more general focus.

> It is a fallacy to infer (à la Superset) from the value-contributing conditions of any given model scenario, that among alternatives that lack at least some of those conditions, supersets of those subsets are better. (2019: 275)

We find much to endorse in this account of the "approximation fallacy". For instance, many discussions of the fallacy tie it to reasoning that approximates to an "ideal" situation. However, Estlund recognizes that one might appeal to approximation whenever it is one's intention to order or rank situations, even when there is no "ideal" situation. But while Estlund's account suitably broadens the scope of reasoning by approximation, the account remains undesirably narrow in several ways.

The first problem is that Estlund's account cannot accommodate instances where approximation is used to establish that a situation with *more of a* value creating condition is preferable over a situation with *less of it*. For example, in the retirement fund case above, with the "discontinuous" policy, there is no uncontrived way of construing the option of contributing $900 as a superset of contributing a lesser sum. Yet, someone who opts to contribute $900, rather than a smaller sum or nothing, because it better approximates a $1000 contribution is committing the approximation fallacy.

Not accommodating "degree-based" value contributing properties, in turn, prevents Estlund's account from capturing approximation-based reasoning when the absence of an attribute increases the value of an option. Suppose you, very plausibly, consider your ideal car to be cheap, safe and in any colour but blue. Now, in choosing between a car that is cheap, safe and indigo (not quite blue, but bluish) and one that is cheap, safe and red, you might prefer the latter because it better resembles (approximates) your ideal car than the former. If you reason in this manner, you are reasoning by approximation. Yet, the set (cheap, safe, red) is not a superset of the set (cheap, safe, indigo).[8]

These issues reflect a deeper problem with Estlund's account. "Approximating" seems to require comparing "objects" in terms of how much they resemble each other (or, how similar they are to each other). The Superset principle, on account of its narrow enumerative stance, obscures the role of such similarity comparisons. Enumeration a-la Superset might (and, in many real life

[8] Substituting "red" with "not-blue" and "indigo" with "bluish" does not work, either.

situations, will) be necessary to establish similarity relations, but it is not sufficient.

Finally, Estlund's approach does not adequately account for legitimate, non-fallacious, uses of approximation reasoning. His account suggests that reasoning by approximation takes the following form:

> P1: Scenario A does not have all of the value creating conditions of the model scenario, yet it has more of them than Scenario B.
>
> P2: Superset principle.
>
> C: Scenario A is better than Scenario B.

This is a valid deductive argument. So, assuming P1 is true, if this argument is fallacious, it must be because premise 2 (which may be implicitly assumed by the arguer) is false. Yet we contend that it is unreasonable to suppose that all instances of reasoning by approximation take this form, and are thereby fallacious. Surely, at least some of those who reason by approximation realize that a scenario that best approximates some model scenario will not always be the next best scenario. In other words, they realize that a principle as general as Superset cannot be relied on. Yet, those who realize this might have good reasons for thinking that, in the domain of interest, approximation to the model scenario is a *reasonable basis* for ranking alternative scenarios.

To accommodate this point, Estlund could supplement the "generalized" Superset principle with "domain-restricted" Superset principles that apply to narrower domains. For at least some domains, the relevant Superset principle would turn out to be true and thus there would be no fallacy in reasoning by approximation in these domains. However, we see two problems with this response.

First, it retains the idea that reasoning by approximation is an exclusively deductive form of reasoning. If the Superset principle applied to a particular domain turns out to be true in most cases in that domain but fails in a small number of cases, then the Superset premise is false and any argument employing it is fallacious. However, in many legitimate instances of reasoning by approximation, although the scenario that best approximates the model scenario is thereby *very likely to be* the next best scenario, there is no guarantee that *it* *will be*. We will give several plausible examples of such cases in the next section. The point for now is that an account of reasoning by approximation should allow for the possibility of ampliative approximation reasoning.

The second problem is that, if reasoning by approximation requires knowing the truth of the relevant (domain-restricted) Superset principle, then it is doubtful that it is often, or ever, employed in real life. This is because in most circumstances where someone knows that a Superset principle is true, they only know this because they have comprehensive knowledge about how intrinsic values and the value creating conditions are related in the relevant domain. But if they have this knowledge, then reasoning by approximation is redundant: they don't need to consider which scenario most resembles the model/ideal/best scenario because their knowledge gives them direct epistemic access to the value in each scenario.

Because of this problem, Räikkä (2014) suggests that in most real life scenarios it is uncharitable to interpret people as reasoning by approximation. Our point is that, although this is a problem for deductive interpretations of approximation reasoning (such as Estlund's), it does not apply to ampliative interpretations. We will elaborate on this further in the next section. For now, we emphasize that our point is not that people could never make deductive approximation-based arguments. It is rather that an account of reasoning by approximation should allow that it comes in both deductive and ampliative forms and that the latter is especially important because realistic instances of approximation reasoning usually take that form.

3. A Better Account of Reasoning by Approximation

Above, we have noted several problems with Estlund's account of reasoning by approximation. These problems suggest to us that a fundamentally different approach is needed. It should follow Estlund by moving beyond an appeal to the "ideal" and the "second-best". However, it needs to go beyond his account by: (1) allowing approximation arguments to be based on differences in the degree to which a value creating condition is present; (2) allowing approximation arguments to explicitly rely on similarity relations; and (3) allowing for ampliative approximation arguments. In this section, we will develop an account of approximation-based reasoning that does all of these things.

We begin by noting that the intellectual objective in any appeal to approximation is to rank “things”, such as objects, choices, and situations. Let us call these things, *targets*. Thus, in Goodin’s example, the cars available for purchase are the *targets* to be ranked. The set of these targets (i.e., the set of “things” that are to be ranked) we will call the “target domain”. As we saw above, in reasoning by approximation, you very often compare hypothetical, but ideal, targets with non-ideal, but actualisable, ones. Thus, the target domain might include hypothetical targets. Notationally, if t_i is the i^{th} target and T is the target domain, we have $t_i \in T$, where $i = 1, \dots, n$ and $i \geq 3$.[9] Each target t_i corresponds *uniquely* to a set of all properties that t_i has. For example, in Goodin’s car, a new, silver Rolls has the properties that it is a car, it is new, it is silver, and it is made by Rolls Royce. So, in our model, a comparison of targets corresponds to a comparison of their sets of properties. Stated formally, $t_i \Leftrightarrow P_i$, where P_i is the set of properties that t_i has.

In addition to a target domain, the task of ranking requires a *basis* for the ranking. We will call this the “value” which the ranking assigns to its targets. This value might be an evaluative property (e.g., ranking cars by how desirable they are, or ranking societies by how just they are). Alternatively, it may be a non-evaluative property (e.g., ranking aeroplane designs by the degree to which they are aerodynamic), although in any realistic example of this kind we would only be interested in the non-evaluative property because we find more or less of it desirable. The ranking function, V, can be expressed as $v_i = V(t_i)$ for $t_i \in T$, where v_i is the value of the target, t_i.[10]

It is also apparent that in assigning a value to a target, the V function must take into account the properties of the targets. For example, suppose that in Goodin’s car, the value of a car is its market value. The function that ranks cars according to market values needs to be a rule that takes into account the car’s make, newness, and “features”, amongst other things. But, V does not have to take into account *all* elements in P_i. Some properties can be neglected because they

[9] Note that $i \geq 3$, and not 2, because appeals to approximation are only meaningful with at least three targets.

[10] Our exposition is silent on whether V is a cardinal or an ordinal function.

are common to all $t_i s$ (such as, $t_i s$ are all cars).[11] Other properties in P_i can be neglected even if they *differ* across t_is because they contribute nothing to the value of the target. For instance, it is likely that the difference in the market value of a *black* Rolls and a *silver* Toyota has nothing to do with the difference in their colours.

Thus far, we have established that, in order to accurately rank targets in a domain (i.e., to assign "v" scores to targets), epistemic access to the V function is required. When we have this access directly, ranking targets by appealing to approximation is redundant. For example, suppose we are ranking, by market value, various gold "coins", "biscuits", and "triangles" produced at a particular foundry. Each of these items corresponds uniquely to a *complete* set of the properties it possesses (e.g., weight, shape, substance of constitution, place of manufacture, etc.). However, the V function will be based on only one property: the weight of each target, w_i.[12] Furthermore, the V function in this instance is a linear function of the target's weight: so, $v_i = V(t_i) = aw_i$; where $a > 0$ and w_i is the weight of t_i.[13] In such circumstances, one would not need to appeal to approximation. One could simply rank the gold coins, biscuits, and triangles by comparing their respective weights.

Unfortunately, even the most quotidian of realistic ranking tasks do not share the simplicity of the preceding example. In these tasks, bounded rationality implies that we do not know what the V function is. Consider, for instance, the task of constructing rankings, in terms of market value, over a target domain consisting of (just) four targets: a large Van Gogh from his late period, a smaller Van Gogh from that period, an in-between sized Van Gogh from his early period,

[11] When producing *rankings* over target domains, one only considers ways in which the targets are *dis*similar. This is analogous to the "Independence of Irrelevant Alternatives" principle found in expected utility theory.

[12] The "independence of irrelevant alternatives" entitles us to disregard substance of constitution and place of manufacture, and differences in shape or time of manufacture are disregarded because we know them to be irrelevant.

[13] A couple of clarifications: first, note that, since our purpose is to produce rankings over the targets, it is not necessary that a represent the market price of gold per unit of weight. Second, if our rankings were ordinal (rather than cardinal), then any monotonically increasing function of target weight (w_i) would suffice. For instance, $v_i = V(t_i) = w_i^2$ would also work.

and a decently sized Vermeer. It is in such instances that one conceivably relies on reasoning by approximation.[14]

The issue, then, is to present the most charitable construal of the structure of such reasoning. We see it as having three main parts. First, the reasoner compares targets in terms of their relevant resemblance to each other (we call resemblance comparisons, "relevant descriptive similarity (RDS) relations"). Second, she uses these "RDS relations" to rank targets in terms of the value she thinks they have (we call these rankings, "RDS rankings"). Third, she uses these "RDS rankings" to infer the (unknown) actual value rankings over the targets (we refer to this as "approximating").

Let's explain the first part in more detail. *Relevant descriptive similarity* (RDS) is, admittedly, a vague concept, but vagueness is an asset in generating a rule of thumb—a heuristic for ranking targets that otherwise, one would not be able to rank. When one does not know *what* the V function is, one might still have (and, indeed, often has) *notions* about which *epistemically accessible* properties of targets are relevant to the V function.[15] For instance, in the Van Gogh/Vermeer example, we know that the market price of an artwork often depends on the artist, size of the work, and period in which the work was produced, yet have an incomplete picture of how these properties "interact" to generate a market price. Having identified the set of properties (R) that are relevant to value, we can consider the conditions under which a target is *plausibly* said to be *more relevantly descriptively similar* than another target to some third target.

The following example teases out some of our basic intuitions on this. Suppose we want to rank societies by how democratic they are and have identified the relevant properties to be:

(i) Holding elections at regular intervals (e)

(ii) People being free to protest against the government (p)

(iii) The *degree* to which the press is free.

[14] In this example, you might be tempted to offer "auctions" as a substitute for approximation. However, "auctions" would not work universally (consider ranking societies in terms of their "moral worth") and, even if an auction were feasible, it would only provide *ex-post* rankings, whereas we are after *ex-ante* ones.

[15] "Epistemically accessible properties" are those that can be easily observed (e.g., colour, shape) or, easily measured (e.g., weight) or, easily assessed (e.g., vintage).

Now, consider the following target societies (in capital letters), compared by these three properties:

A: $(e; p;$ press is free to the degree $x)$
B: $(e; p;$ press is free to the degree $y < x)$
C: $(e; p;$ press is free to the degree $z < y)$
D: (not $e; p;$ press is free to the degree $y)$
E: $(e;$ not $p;$ press is free to the degree $y)$

Given this, the following claims are each intuitively compelling:

(1) A is *more* relevantly descriptively similar to B than A is to C;

(2) B is *more* relevantly descriptively similar to A than B is to C, iff $|x - y| < |y - z|$;

(3) A is *more* relevantly descriptively similar to B than A is to D;

(4) Nothing can be said about whether B is more relevantly descriptively similar to A rather than to D;

(5) There can be no comparisons of relevant descriptive similarity between B, D and E.

These claims stem from the basic assumptions that we rely on in comparing objects for their relevant descriptive similarities. The following heuristic captures these assumptions:

Heuristic 1 (for establishing RDS relations):

For any triple of targets, t_i, t_j *and* t_k, and a relevant (similarity-contributing) set of properties, R, t_i is more relevantly descriptively similar to t_j, than it is to t_k iff:

i) For any binary property t_k has in R_k, t_i and t_j have it in R_i and R_j (and, for any degree-based property that t_k has in R_k, t_i and t_j have it in R_i and R_j to, at least, the same degree as t_k has in R_k); and,

ii) There is some binary property t_i and t_j have in R_i and R_j that t_k does not have in R_k (or, there is some degree-based property t_i and t_j have in R_i and R_j to a greater degree than t_k has in R_k).

Heuristic 1 helps us to establish RDS relations between targets. What our account now needs is an understanding of *how* RDS relations translate to an RDS-based ranking over the targets. Similarity relations alone cannot produce an RDS-based ranking of targets; we also need some way of assessing how the similarity we observe is related to the value we want to track. We see two different ways of making this assessment which lead to two (slightly) different methods of reasoning by approximation. In principle, each method can be applied to the same cases. However, in practice, the epistemic limitations you are under will usually make one or the other method more suitable and hence dictate which form of approximation reasoning you would use.

The first method (Method 1) is used in cases where you have no way of conclusively determining which target, actual or hypothetical, is, in fact, the most valuable (i.e., you have no model/ideal target to approximate to). In these cases, to rank available targets by their similarity relations, you must use your beliefs about how the values of targets depend on *each individual* property in the set of RDS properties.

For example, suppose that at a recent telehealth consultation your doctor diagnosed you with two minor unrelated medical conditions. A few days later, a package arrives from your doctor containing two different bottles of pills. You have the dosage for these pills but lose the instructions about how they are optimally used. In this situation, although you do not know at the outset which target (way of using the medications) is the most valuable, you still can construct similarity relations between the targets, S (use both prescribed medication), T (use only one) and, U (use none). To go from similarity relations to similarity-based orderings, you would need to invoke your "background" beliefs about the world. You might reason that, since unrelated conditions need distinct treatments, taking *each* pill contributes, independently, to the value in a situation. Such reasoning, together with the similarity relations, will then lead you to rank S over T over U.

Before suggesting a general heuristic for producing similarity-based rankings, let's test our intuitions with the five societies (A, B, C, D and E) compared above. Our background knowledge tells us that, other things being equal, both holding regular elections and permitting the articulation of dissent will independently increase democratization; it also tells us that, other things being equal, the higher the degree of press freedom, the higher the degree of democratization. Heuristic 1 tells us that A is more RDS-similar to B than it is to C. So, given background beliefs and given similarities, we would *intuitively judge* A to be more democratic than B (in other words, B is inferior to A) and B to be more democratic than C (in other words B is superior to C). Finally, without additional beliefs about how each property is to be weighted, we wouldn't intuitively infer any other similarity-based ranking over the five target societies. The intuitions appealed to in the above discussion are captured in the following heuristic.

> *Heuristic 2* (for establishing RDS rankings)
> t_j is RDS-inferior to t_i and RDS-superior to t_k (for expositional simplicity, we can state this as $RDS_i > RDS_j > RDS_k$) iff:
>
> i. The value of a target increases when each (binary-valued) property identified as relevant is present (rather than absent)
> ii. The value of the target is increasing in the level of each (continuous-valued) property identified as relevant
> iii. t_i is more relevantly descriptively similar to t_j than t_k[16]

As we have seen, Heuristic 2 is used in approximation reasoning that follows what we have above called "Method 1". However, many instances of approximation reasoning follow a different method that identifies an actual or hypothetical target with the maximum value and uses similarity to this target to generate RDS rankings. This method, "Method 2", can be illustrated with the following example. Suppose your doctor has prescribed you two pills, to be taken daily. She has been vague otherwise (about what your diagnosis is, or what the pills are for). As before, your options are S (use both prescribed medications), T (use only one) and, U (use none). Given your doctor's expertise, you conclude that S is the most valuable target. Furthermore, your background beliefs (let's suppose) suggest that only the number of pills taken is relevant to the value of the target. Given this, you can use Heuristic 1 to infer that S is more relevantly descriptively similar

[16] If there are properties that are intuitively value-reducing they need to be re-interpreted negatively before being used with this *Heuristic 1* (for instance, the degree of electoral violence would be reinterpreted as the degree to which elections are peaceful).

to T than it is to U. Given that S has the most value of all targets, you can then infer that $RDS_S > RDS_T > RDS_U$.[17]

A few clarifications are in order here. First, if the relevant set of properties consists exclusively of binary-valued properties whose presence increases v values, then RDS-rankings over targets *could* be based on—and this is consistent with our account—satisfying more (rather than, fewer) of these properties. In this case, RDS-rankings will be consistent with Estlund's Superset principle. For example, if we are ranking cars in terms of their safety and we know that the presence of certain features (e.g., airbags, anti-lock braking, etc.) increase a car's degree of safeness (and the lack of such features subtracts from it), then a car will be RDS-superior to another if the former's features are a *superset* of the latter's features (i.e., the former has all the features of the latter and at least one additional feature).

Second, it is possible that the relevant set of properties is singleton and contains an element that comes in degrees. If v values are an increasing function of the value of this element, then, RDS-rankings over targets will be based on *how much* of the property each target has. For example, if we are ranking societies by their degree of income inequality, and if the relevant set contains the value of the Gini coefficient (a standard measure of income inequality), then the RDS ranking of each society will correspond to its Gini coefficient ranking.

Third, it is possible that the relevant set of properties is, *prima facie*, singleton and contains an element that cannot be meaningfully represented as coming in degrees. For example, suppose that we are ranking aeroplane designs by the degree to which they are aerodynamic. Suppose also that "aerodynamism" depends only on the shape of the design. In that case, many will find it intuitive that plane A has a higher RDS rank than plane B, if and only if A has a shape that resembles (i.e., is more relevantly descriptively similar to) the most aerodynamic

[17] So, regardless of what method is used, our account demonstrates that any genuine attempt at reasoning by approximation must rank targets by the properties they possess. This is an important insight. While agents are driven to reason by approximation when they have no direct access to the "value" of the targets, in order to do so, they must: (1) identify value-relevant properties (argued for earlier); (2) use the distribution of those properties in the targets to assess degrees of similarity between targets (argued for earlier); and, (3) assess how each value-relevant property individually contributes to value (argued here).

shape.[18] This might, at first glance, appear to spell trouble for our account because *Heuristic 1* cannot work with a property such as shape. But being irregularly shaped is reducible to several distinct properties (e.g., number of angles, the gradient of curves, etc.) and this reduction is what we quite plausibly resort to in comparing irregular shapes. Once this reduction is done, *Heuristic 1* is up and running again.

Finally, one could object that *Heuristic 1* is unable to accommodate the following example. Suppose that your ideal salad is made from walnuts, Italian dressing, radicchio, and ricotta cheese. Any chef will tell you that a salad that substitutes Belgian endive for the radicchio and cottage cheese for the ricotta cheese will better approximate the taste of this ideal salad than a salad made from only walnuts, Italian dressing, and radicchio. Yet *Heuristic 1* seems unable to account for this. Our response is that, when this example is accurately described, *Heuristic 1* does account for it. For if the basis of comparison is overall taste, then the ideal salad should be understood in terms of a "taste profile" rather than in terms of specific ingredients. In other words, rather than saying that your ideal salad contains ricotta cheese, we should say that it contains a cheese with the set of relevant "taste profile" properties that ricotta has. By adding this detail, we get a better explanation of why the salad with the two substitute ingredients better approximates the ideal salad. Cottage cheese is not an exact match for the taste profile of ricotta, yet it shares many of its taste profile features (e.g., creamy curds), and, when a feature comes in degrees, their degrees are often close (e.g., cottage cheese is only slightly less sweet than ricotta). Given this, the salad that uses Belgium endive and cottage cheese will match the various taste profiles of the ideal salad to a greater degree than the salad that uses radicchio (a perfect match) but has no cheese. But once the example is understood in these terms, it is consistent with *Heuristic 1*.

Given our account of what it means for a target to be more relevantly descriptively similar to another, rather than to a third, and given our account of how targets are ranked by their relevant descriptive similarities, we are now ready to give the general form of arguments that use approximation reasoning.

[18] The reader will notice that this is another "Method 2" example.

The Argument by Approximation

1. The relevant set of properties is R.
2. t_1 is more RDS to t_2, than it is to t_3 and, t_2 is more RDS to t_3, than it is to t_4, and so on. [1, *Heuristic 1*]
3. The RDS-ranking over targets is: $RDS_1 > RDS_2 > \cdots > RDS_n$. [2, *Heuristic 2*]
4. For any t_i, t_j, t_k, $Prob(v_i > v_j > v_k | RDS_i > RDS_j > RDS_k) = \alpha$.
5. Therefore: $v_1 > v_2 > \cdots > v_n$. [3,4, Approximation][19]

Premise 4 is crucial to reasoning by approximation. The value of α in this premise reflects the degree of "correspondence" that the arguer takes there to be between the proxy RDS ranking and the actual value ranking. This directly addresses the question of whether an argument by approximation is best characterized as deductive or ampliative. When $\alpha = 1$, the inference is valid and, the argument is deductive. When α is less than 1, yet reasonably high, an ampliative interpretation of the argument is most plausible. Generally, the value of α will be relatively high when there is a monotonic relationship between values of properties and values of targets and no "interaction" effects are present (i.e., in circumstances where RDS rankings are correctly inferable). This is consistent with views in the literature (discussed in §1) about the circumstances in which approximation reasoning is fallacious.

Three examples, further demonstrate these differences. First, let us return to the task of ranking gold coins, triangles, and biscuits in terms of their market value. Recall that, in this example, market value depends positively *only* on the weight of the targets. Now, if weight is used to generate RDS-rankings over these objects, then these rankings will "perfectly correspond" to rankings in terms of

[19] This argument form corresponds to what we above called "Method 1". When an argument by approximation follows Method 2 an additional premise stating that " t_1 is the model target" is added. This additional premise, along with premise (2), is then used to derive premise (3), without appealing to *Heuristic 2*.

market value (i.e., α is equal to 1). Hence, if you used approximation reasoning in this case, it would be deductive.[20]

As a second example, imagine that you want to rank pieces of gold-only jewellery by their market value, using the weight of each piece as the RDS-ranking. Common sense tells us that, although each piece's weight significantly and positively contributes to its market value, variations in weight can be neutralised by counteracting variations in craftsmanship. Thus, there would be an imperfect, yet still significant, "correspondence" between the jewellery's RDS-rankings and comparative market values (i.e., α will be less than 1, yet still relatively high). Hence, the argument by approximation used in this example is best understood as a relatively strong ampliative argument.

For the final example, suppose that you are ranking (again, in terms of market value) a newly minted gold coin, an 1837 gold sovereign, an intricately designed contemporary gold necklace, and a plain gold bangle that Princess Diana wore during the fatal paparazzi chase. Weight would not be a significant factor in the comparative market value of these items. Thus, a weight-based RDS-rankings will not only "imperfectly correspond" to market value-rankings, its correspondence will also be extremely weak (i.e., the value of α will be low). Thus, any approximation argument made here would be very weak. This is probably why it is inconceivable that one would reason by approximation in this case.

Given the preceding discussion, it is apparent to us that a particular merit of our account of approximation reasoning is that it allows such reasoning to be either deductive or ampliative. The possibility of ampliative approximation arguments allows one to evaluate—in terms of strength—appeals to approximation made in different contexts. This is an improvement over a strictly deductive interpretation, which doesn't seem to accurately capture many cases of approximation reasoning and sometimes misclassifies legitimate approximation reasoning as fallacious

A second advantage of our approach is that it provides handy guidance on the contexts in which approximation reasoning is likely to succeed and likely to fail. To see how, let us return to the discussion on how RDS-rankings are

[20] Of course, in these simple sorts of examples, reasoning by approximation is redundant, since as argued before, the V function is known.

generated. The heuristic (*Heuristic 2*) used to produce RDS-rankings requires *each* element in R_i to *individually* contribute *at least as much* to value as the corresponding element in R_j. When such rankings can be constructed, reasoning by approximation is more likely to be viable. This is consistent with our intuitions in the cases where approximating leads us astray. Consider, again, the example of Estlund's pills: to use approximation-based reasoning according to our account, one would have to claim that taking only one of the pills is RDS-superior to taking no pills and RDS-inferior to taking both pills. But our RDS-ranking heuristic (*Heuristic 2*) precludes such a claim where interaction effects between pills are present! It is *not the case*, in this instance, that the presence of *a* pill, by itself, contributes to value; rather value is *jointly* generated by the *pills*. Once this is acknowledged, it becomes apparent that if one insists on reasoning by approximation—despite the patent unsuitability of the situation for generating RDS rankings—one will, most likely, end up reasoning fallaciously. However, if there were no interaction effects (or, if your background beliefs were correct), the RDS ranking would be accurate, and reasoning by approximation would be appropriate.

Another example commonly discussed in the literature is the "fallacy" of inferring that an economy with more competitive markets is more efficient than one with fewer competitive markets because it is descriptively more similar to the most efficient economy, where all markets are competitive.[21] Our account shows why approximation-based reasoning is inappropriate here. It is inappropriate because RDS-rankings cannot be produced in this context. To produce such rankings using the methods we highlight above, it would have to be that an individual market's competitiveness, *ceteris paribus*, will increase efficiency in the economy. But, as Lipsey and Lancaster showed, that might not be the case, since prices of goods are interdependent, and changing the price of one, and leaving others unchanged, might misallocate resources across markets.

Reasoning by approximation is also inappropriate when RDS-rankings cannot be produced because RDS relations cannot be meaningfully established between targets. For example, approximating would be futile when comparing the market value of gold objects with different weights, degrees of inherent

[21] See, especially Räikkä (2000) and Heath (2010), Chapter 3.

craftsmanship, vintages, and idiosyncratic characteristics. This is because, despite the objects all being made of gold, there is no plausibly meaningful way of establishing which object is more similar to another than it is to a third. Using any one of the value creating characteristics (say, vintage) will give rise to similarity relations that are inconsistent with those resulting from using another characteristic (say, craftsmanship). In other words, *Heuristic 1* cannot be used.

Finally, when reasoning by approximation is fallacious, our account has the advantage of identifying where exactly the problem lies. Let us illustrate this with Goodin's car. Suppose your targets are a new silver Rolls, a one-year-old black Mercedes and a new silver Ford, and you are ranking these targets according to their desirability (v). Suppose, also, that v depends, for you, on three characteristics: (N)ewness; (L)uxuriousness; and (C)olour.[22] So, the value in any target, t, is given by $v_t = V(N^t, L^t, C^t)$. If the explicit form of the V function is not known to you, you might reason by approximation to derive a preference ordering of the cars. Goodin argues that if someone infers that the Ford is more desirable than the Mercedes, given that the ideal car is a new silver Rolls, they would be fallaciously reasoning by approximation. But why is this fallacious?

By fitting Goodin's example into our argument form, we can show exactly where the problem lies. To do this, we will use the following additional notation for expositional ease. Recall that the Rolls is more luxurious than the Mercedes which is more luxurious than the Ford: using $L^{\mathrm{F}}, L^{\mathrm{M}}, L^{\mathrm{R}}$ as the degrees of luxuriousness of the Ford, Merc and Rolls, respectively, we have $L^{\mathrm{F}} < L^{\mathrm{M}} < L^{\mathrm{R}}$. Also recall that both the Rolls and the Ford are brand new, while the Mercedes is one year old: representing $N^{\mathrm{F}}, N^{\mathrm{M}}, N^{\mathrm{R}}$ as the newness of the cars, respectively, we have $N^{\mathrm{M}} < N^{\mathrm{F}} = N^{\mathrm{R}}$. Finally, recall that while, the Ford and the Rolls are both (desirably) silver, the Mercedes is black: using $C^{\mathrm{F}}, C^{\mathrm{M}}, C^{\mathrm{R}}$ as the colour of the cars, respectively, we have $C^{\mathrm{F}} = C^{\mathrm{R}} = (S)ilver$ and $C^{\mathrm{M}} = not\ S$. Given this, the argument can be represented as:

1. $R = (N, L, C)$

2. $Rolls = (N^R, L^R, C^R = S)$ & $Merc = (N^M, L^M, C^M = not\ S)$ & $Ford = (N^F, L^F, C^F = S)$

[22] N and L are continuous-valued properties while C is binary-valued.

3. $Ford$ is more relevantly descriptively similar to $Rolls$ than $Merc$ is to $Rolls$ [2, *Heuristic 1*]
4. $Rolls$ is the ideal target
5. $RDS_{Rolls} > RDS_{Ford} > RDS_{Merc}$ [3, 4]
6. $Prob(v_{Rolls} > v_{Ford} > v_{Merc} | RDS_{Rolls} > RDS_{Ford} > RDS_{Merc}) = \alpha$
7. Therefore: $v_{Rolls} > v_{Ford} > v_{Merc}$ [5, 6, Approximation]

Now we see exactly where the problem is: 3 is false and misinferred from 2 because *Heuristic 1*, by which relevant descriptive similarity between targets is established, is misapplied; the Ford is not equally, or more, luxurious than the Merc. Therefore, the problem is neither in inferring the RDS-rankings, nor in inferring, from those rankings, personal preferences over targets; the problem is with a false premise. Note also that, if the Ford were to be replaced by a one-year-old, silver Jaguar (a more luxurious car than a Merc) you would have strong approximation-based grounds to prefer the Jaguar over the Mercedes. What then can we say to someone who needs to choose between the Ford and the Mercedes, wants to use approximation reasoning, and values the luxuriousness of the car more than other attributes? In such a scenario, our advice would be to amend the set R. That is the sensible thing to do, since no available option satisfies all three criteria. With $R = (L)$, following the structure as before, you would come to the (intuitive) conclusion that the Mercedes (and not the Ford) is the car to buy.

To summarise, our account shows that an argument by approximation can go wrong if any of the following occurs: (1) R is incorrectly specified, (2) the attributes of the targets are incorrectly identified, (3) relevant descriptive similarity relations between targets are incorrectly established, (4) RDS rankings are misinferred, (5) RDS rankings have a weak correspondence to v rankings. Bar these failings, there is nothing, in principle, fallacious about reasoning by approximation. This is a more nuanced account of the "fallacy" of approximation than the simple treatment it is often given in the literature. Indeed this account may lead some to question whether the errors that can occur in approximation reasoning are sufficiently common, systematic, and unified to warrant them being described as a "fallacy".

In conclusion, we suggest that our account has put appeals to "reasoning by approximation" and the "fallacy of approximation" on a more secure footing.

It avoids the various problems that beset other accounts, and captures all the contexts in which approximation reasoning occurs. It also shows how such reasoning can be either deductive or ampliative. This is important as recognizing that approximation reasoning has an ampliative form shows that it is more credible than is often supposed. Finally, our account gives a nuanced breakdown of the various ways in which reasoning by approximation can go wrong. As a result, philosophical debates where approximation reasoning plays a role can now proceed with more care and precision.

Acknowledgements: We would like to thank three anonymous referees from this journal for providing astute critical feedback on this article that helped us to better approximate the ideal we were aiming at. We are also grateful to Daniel Nolan for helpful comments on an earlier draft of this article.

References

Berg, A (2019). “Incomplete Ideal Theory”, *Social Theory and Practice*, 45 501-524.

Brennan, G. and P. Pettit (2005). “The Feasibility Issue”. In F. Jackson and M. Smith (Eds.), *Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Brennan, G. (2007). “Economics”. In R.E. Goodin, P. Pettit, and T. Pogge (Eds.), *A Companion to Contemporary Political Philosophy*, Wiley-Blackwell.

Cavallero, E. (2011). “Health, Luck and Moral Fallacies of the Second Best”. *Journal of Ethics*, 15: 387-403.

Coram, B. (1996). “Second Best Theories and the Implications for Institutional Design”. In R.E. Goodin (Ed.), *The Theory of Institutional Design*. New York: Cambridge University Press.

Elster, J. (1986) “The Market and the Forum: Three Varieties of Political Theory”. In J. Elster and A. Hylland (Eds.), *Foundations of Social Choice Theory*, Cambridge: Cambridge University Press.

Estlund, D. (2019). *Utopophobia: On the Limits (If Any) of Political Philosophy*. Princeton: Princeton University Press.

Goodin, R.E. (1995). “Political Ideals and Political Practice”. *British Journal of Political Science* 25: 37–56.

Goodin R.E. (2012). “The Bioethics of Second-Best”. In J. Millum and E. Emanuel (Eds.), *Global Justice and Bioethics*, Oxford: Oxford University Press.

Heath, J. (2010). *Economics without Illusions: Debunking the Myths of Modern Capitalism*. New York: Random House.

Heath, J. (2013). “Ideal Theory in the *n*th Best World: The Case of Pauper Labor”. *Journal of Global Ethics* 9: 159-172.

Lipsey, R. and K. Lancaster (1956). “The General Theory of Second Best”. *The Review of Economic Studies* 24: 11–32.

Margalit, A. (2010). *On Compromise and Rotten Compromises*. Princeton: Princeton University Press.

Ng, Y.K. (2004) *Welfare Economics*. New York: Palgrave McMillian

Pettit, P. (2006). “Can Contract Theory Ground Morality?” In J. Dreier (Ed.), *Contemporary Debates in Moral Theory*, Wiley-Blackwell.

Räikkä, J. (2000). “The Problem of the Second Best: Conceptual issues”. *Utilitas* 12: 204–18.

Räikkä, J. (2014). *Social Justice in Practice*. Switzerland: Springer.

Sen, A. (2009). *The Idea of Justice*. Cambridge, MA: Harvard University Press.

Steinberg, E. (2017). “The Inapplicability of the Market-Failure Approach in a Non-Ideal World”. *Business Ethics Journal Review* 5: 28-34.

Wiens, D. (2016). “Assessing Ideal Theories: Lessons from the Theory of Second Best”. *Politics, Philosophy and Economics* 15: 132–49.

Wiens, D. (2020). “The General Theory of Second Best is more General than you think”. *Philosophers Imprint* 20: 1-26.