Ethics and Information Technology (2023) 25:47
https://doi.org/10.1007/s10676-023-09722-w

ORIGINAL PAPER

Digitising reflective equilibrium

Charlie Harry Smith[1]

Abstract
Reflective equilibrium is overdue a twenty-first century update. Despite its apparent popularity, there is scant evidence that theorists ever thoroughly implement the method, and fewer still openly and transparently publish their attempts to do so in print—stymying its supposed justificatory value. This paper proposes digitising reflective equilibrium as a solution. Inspired by the global open science movement, it advocates for coupling a novel, digital implementation of the equilibrating process with new publication norms that can capitalise on the inherent reproducibility of digital data. The argument is structured around three main claims: that digitising will make it easier to (a) methodically construct, (b) widely disseminate, and (c) thoroughly critique reflective equilibria. Practical guidance is also provided throughout. Altogether, it is argued that digitisation will not only help theorists to better realise reflective equilibrium's latent theoretical potential, but also greatly extend its value as a justificatory device in contemporary academic discourses.

Keywords Reflective equilibrium · Digital reflective equilibria · Digital humanities · Philosophical methods · Open science · Reproducibility

Introduction

Popularised by Rawls (1951, 1974, 2005a), and thus forever guaranteeing its favour amongst liberal theorists, reflective equilibrium (RE) is generally considered the leading methodological tool in contemporary political philosophy (List & Valentini, 2016, p. 542). It has even been called "the" method of philosophy *tout court* (Lewis, 1983, p. x). Yet RE is well overdue a twenty-first century update, due to two interrelated issues:

i. Most accounts do not set theorists up for success when it comes to actually implementing the method (cf., Rechnitzer, 2022), seriously underplaying just how difficult and time-consuming it can be to construct thoroughgoing reflective equilibria using traditional tools.
ii. Prevailing research dissemination norms needlessly conceal the method's primary output—the equilibria in question—from a theorist's readers, thereby shielding many of their argumentative commitments from critical scrutiny.

The second issue is more damning than the first. Even assuming that entire reflective equilibria are being constructed somewhere behind the scenes, the decision not to share them means readers must simply trust that a coherent whole does in fact exist—either in theorists' heads or their notes. This needlessly inconveniences readers, effectively requiring them to speculate and infer what the details of that wider whole might be. After all, they can do little more than extrapolate from any limited fragments of an equilibrium that might happen to make it into a theorist's essays, chapters, and monographs. But in a discipline that claims to pride itself on argumentative rigour and clarity, such appeals to blind trust or guesswork should feel distinctly out of place. Indeed, I am surprised that we seem happy to accept this status quo. It drastically limits critics' opportunities to scrutinise equilibria in their entirety, leaves plenty of room for misunderstanding, and so significantly hinders RE's supposed justificatory value.

For RE to truly earn its preeminent status, its supporters need to reassert its value in the face of these problems. In this paper, I therefore attempt to do precisely that. My suggestion is that certain computer programs have much

✉ Charlie Harry Smith
charlie.smith@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, United Kingdom

Published online: 07 September 2023

to offer theorists committed to realising RE in practice. Primarily, they provide those of us working in the age of networked computing with better ways to construct and share our philosophical arguments. I accordingly advocate for the adoption of a novel, digital implementation of RE, coupled with some quite radical publication norms that take seriously the inherent reproducibility of digital data. In advancing this argument, I make three main claims: that such digitisation will make it easier to (a) methodically construct, (b) widely disseminate, and (c) thoroughly critique reflective equilibria.

Structurally, the paper expands on each of these claims in turn. After further detailing the implementation issues identified at the outset, section one shows how digital tools can help theorists to capture the complex interconnectivity and inherent non-linearity of the RE process. My argument is that adopting these tools is likely to enhance the general clarity and systematicity of RE-based theorising. Then, when it comes to dissemination, I propose in section two that working digitally would benefit both theorists and their readers by opening up new avenues for sharing and publishing reflective equilibria online—ideally via agreed upon, open formats—to bolster argumentative transparency, flexibility, and explainability. In the third section, I contend this could support new practices around the auditing and peer-review of entire equilibria, furthering the method's rigour. Finally, and more speculatively, I suggest that working digitally could even enable radically collaborative applications of RE.

Of course, there will also be downsides to digitisation. Tracking the development and spread of ideas is a particular challenge for current digital solutions. If left unaddressed, this could make it hard to see how positions have evolved over time, as well as raising possible intellectual property issues. I therefore conclude by suggesting enhancements that would make existing computer programs better suited to realising the method. But, for the theorist interested in deploying RE here and now, this paper offers pragmatic advice for applying a user-friendly, digital implementation of the method in practice. Whilst it transpires that the thorough application of RE is still time-consuming, even with the help of digital tools, I nonetheless affirm that digitising the process has real methodological value. Digitisation promises to not only help theorists better realise RE's latent theoretical potential in practice, but also extend its value as a justificatory device in normative discourses.

Constructing equilibria

My first major claim is that certain digital tools can capture the inherent complexity and non-linearity of the RE process more effectively than traditional approaches. By this, I mean that theorists working digitally are more likely to successfully achieve the construction of rigorous and complete equilibria. Further, they are likely to find the digital RE construction process more efficient. To help illustrate these claims, I will begin by sketching the received method, which will serve as a continued point of comparison.

In a normative[1] context, the RE theorist's basic aim is to bring their "considered judgements" about a moral and/or political situation into a state of coherence with a set of principles they develop to systematise them (Rawls, 1974, p. 8). This state is to be achieved via a process of "mutual adjustment" between those judgements and principles (Rawls, 2005a, p. 20), and results when the theorist manages to minimise any conflicts between these elements whilst also maximising their supportive interrelations (Cath, 2016, p. 216). If, after sufficient revision across levels, they manage to find a coherent and consistent whole, we say the theorist holds their beliefs in a state of reflective equilibrium. We call it (a) an *equilibrium* because all the various elements offer each other mutual support and fully cohere, and (b) *reflective* because the theorist has adjusted these elements without prejudice, revising across all levels of abstraction as they moved back and forth (Doorn & Taebi, 2018, p. 492; Rawls, 2005a, p. 18).

Stopping here, however, would leave the theorist with only a 'narrow' RE (Rawls, 1974, p. 8). A 'wide' RE, possessing greater justificatory power, can be reached by further setting these judgments and principles against various moral and non-moral background theories. Coherence is now to be sought via mutual adjustment across all three of these levels (Daniels, 1979, p. 259). Note that these theories are more than mere "reformulations" of our emergent principles (Daniels, 1979, p. 260); they instead provide their relevant "normative background" (De Vries & Van Leeuwen, 2009, p. 491). In other words, they constitute different sets of comprehensive moral and political doctrines, such as utilitarian or republican approaches, as well as any normatively salient theories about how the world works[2]. Rawls, for instance,

[1] Although my focus here is on normative applications of RE, I suspect digitisation should also benefit projects applying the method in other contexts—like those of, e.g., Elgin (1999) and Lewis (1983).

[2] There is disagreement about what counts as a background theory—or indeed as a principle or considered judgement—but the specific content of each element is not relevant here. What is relevant is just how much theorists are supposed to be systematising via the ongoing process of mutual adjustment. See, for instance, the extensive list of consideration that Daniels (1979, pp. 338–339) suggests.

includes a specific theory of personhood along with descriptive claims about the nature of society.

Realising reflective equilibria

It is when we consider what balancing all these considerations would practically entail that the trouble starts. Theorists should be charting all the possible interconnections between an extensive set of judgements, a further set of ever-developing principles, and numerous background theories, constantly revisiting and revising all of these elements as they go. They will presumably also need to reliably record what combinations they have already tried, why different elements were accepted or rejected, as well as which combinations remain up for evaluation. If that is not already enough, this complexity will only be compounded by the need to repeatedly weigh elements against each other in a cyclical manner, as any adjustments in the system may always have knock-on effects elsewhere.

As anyone who has ever tried will attest, it therefore takes an awful lot of work to reach a stable RE. That is, if reaching one is achievable at all—Rawls (2005, p. 97) famously thought "the struggle for reflective equilibrium continues indefinitely", as there would always be more content to integrate down the line. Yet despite how "demanding" or potentially even "unachievable" finding an RE is, it is nevertheless considered the ideal that theorists should be striving for (Knight, 2017, pp. 49–50). You quickly begin to wonder if this is why "few real applications of the method are presented in the literature" (Doorn & Taebi, 2018, p. 508). Time-poor, modern academics simply cannot commit years of their lives to crafting the sprawling, complex equilibrium that could underpin their own *A Theory of Justice*. Indeed, I suspect few theorists ever actually follow the RE process to completion, which perhaps explains why even fewer have thoroughly, openly, and transparently published their attempts to do so in print[3].

One way to ease the burden is of course through the employment of external aids. Perhaps, in neatly circumscribed situations, some exceptionally talented individuals may be capable of constructing narrow reflective equilibria within their own heads (it is certainly not possible for me). But I doubt anyone could reach a wide RE without technical assistance. At the very least, pens and paper, whiteboards, spreadsheets, or word processors are going to be required to record (a) which judgements have been selected or rejected, (b) which principles have been tried, tweaked or tossed aside, (c) how background theories might speak for or against the developing equilibrium, and (d) how all these considerations will interact. Evidently, tracking all these complex interconnections with traditional tools will be exceedingly unwieldy and time consuming.

The problem is that academic philosophical writing is decidedly linear. Our arguments generally unfold via continuous narrative threads, so the writing tools and approaches we are taught to employ are geared towards this linearity. But serialised sentences and paragraphs—recorded in reams of notes and lengthy Word documents—are poorly suited to capturing an ever-evolving web of interconnected judgements and principles (Fig. 1). There is no one starting point to an RE, with many possible routes into and through the mutually supportive web available. Even analogue tools supporting a more freeform approach cannot effectively handle the process's sheer scale, with all its potential for rewrites and refactoring. The task is vast, multi-dimensional, and recursive, with each element usually offering holistic support to many of its neighbours. In short, there is a real practical challenge here for the theorist attempting to reach an RE with the help of traditional tools, rendering them ineffectual for realising the method.

To be clear: I am not suggesting to do away with the essay or monograph. Translating the circuitous back and forth of RE into linear form is certainly possible and, later in the process, even highly desirable. But it is important to recognise that it is far easier and more appropriate to do this towards the end of a project, *when* and *if* a coherent equilibrium that can be neatly summarised has been found. The theorist's task then shifts: how can they best convey the most important parts of the RE they have constructed to their readers? Before then, though, it makes far more sense to work non-linearly, in a way that reflects the distributed structure of the RE process itself. Thankfully, digital tools supporting such an approach already exist. Although they were not developed for RE, they nevertheless offer features that can be readily adapted for its purposes.

Repurposing zettelkasten software

Whilst mind-mapping programs might immediately spring to mind when faced with a non-linear problem space, I will instead focus on a subset of note-taking applications designed for 'personal knowledge management'. My chief recommendations are Zettlr (Erz, 2023), Logseq (Numerous Authors, 2023), and Obsidian (Li & Xu, 2023)[4]. These

[3] As Rechnitzer (2022, p. 1) notes, the few examples that do exist in the literature "either use a sketchy conception of RE, are restricted to simplified cases, focus only on some particular elements of RE, or do not make the application explicit and traceable." Her excellent book is a notable exception.

[4] I have recommended free applications for accessibility reasons, and prefer open-source tools. See the below Section on 'Open Scholarship' and (Erz, 2019) for why. Nonetheless, other programs, including commercial options, may also support the construction of DRE. Plugins for text editors like Emacs and Vim, for instance, can bring

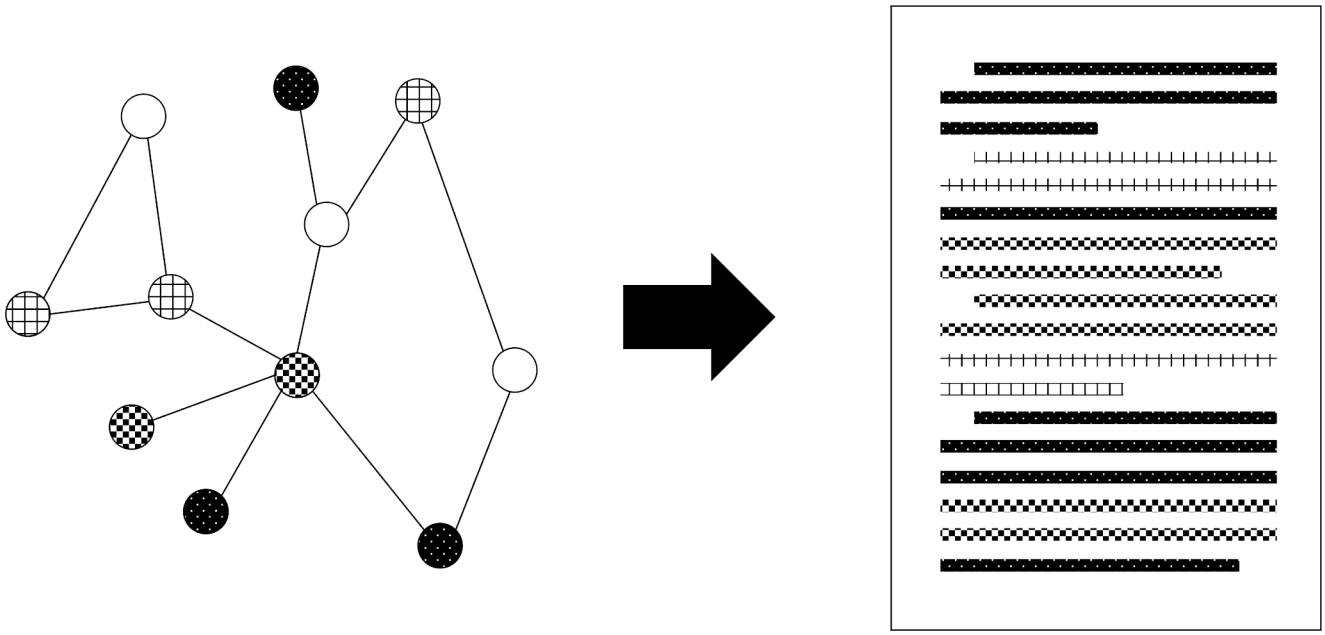

Fig. 1 A diagram representing a distributed web of ideas being translated into a linear essay structure

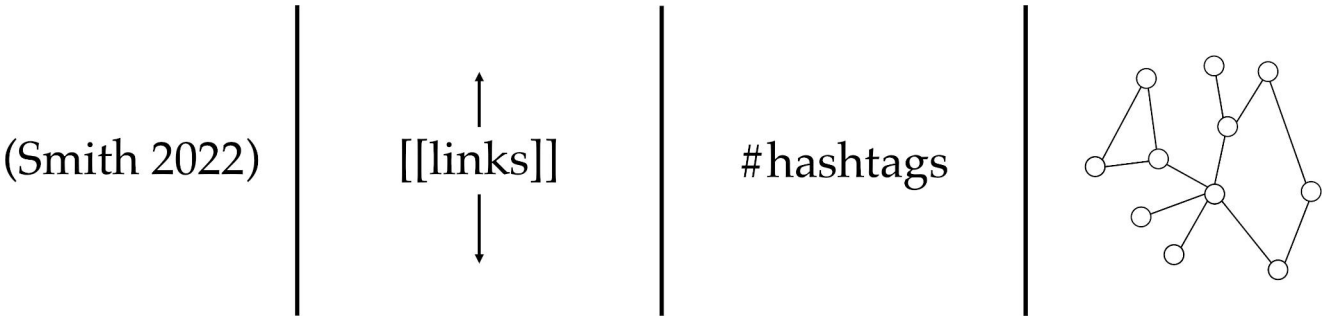

Fig. 2 Diagrammatic examples of the primary application features relevant for realising DRE; inline bibliographic citations, bidirectional hypertext links, hashtags, and graph visualisations

'zettelkasten' applications, which draw inspiration from the slipbox of notes maintained by prolific German sociologist Niklas Luhmann, possess toolsets that make them far more powerful than mind-mapping programs. As both are also free, enjoy committed online followings and operate on what are essentially plain-text 'markdown' files, they represent relatively safe and accessible options for modern academics. Online support networks are extensive and any basic text editor can open the files they create, so research data should remain accessible even if these particular programs are one day depreciated and replaced.

Repurposed for our ends, zettelkasten programs can empower theorists to build a database of notes that connect together in ways not supported by traditional note taking solutions, readily mimicking the structure of an RE. Of course, they include all the basic writing, editing, and formatting functions you would expect. In addition, both also support relatively unique features which make them particularly well-suited to realising the RE process. These are: inline bibliographic citations[5], bidirectional hypertext links, hashtags, and graph visualisations (Fig. 2). Whilst only the first two features are perhaps essential for constructing what I will henceforth call *digital reflective equilibria* (DRE), the latter two result in significant quality of life improvements for theorists and their readers. Altogether, I will argue they push the utility of DRE far beyond their analogue ancestors.

Let me say a little more about these features. The first, *citation support*, is all but required in academic writing. Many scholars already use reference managers (like Zotero, Endnote or Mendeley) to track their sources, and Zettlr,

them to feature parity with zettelkasten software. So-called 'argument mapping' tools also exist but, in my view, do not support sufficient functionality for constructing wide DRE, making them better suited to mapping smaller arguments within an RE.

[5] Zettlr and Logseq support this functionality out of the box. Obsidian requires a citations plugin to be installed.

Logseq, and Obsidian can all integrate with these applications to ease the referencing process during writing. Bibliographies and in-text citations can then be handled in the background, allowing theorists to track where they are appealing to for textual support and so ensure that they integrate others' ideas without risking plagiarism. As the RE process usually deals with a wide range of interconnected arguments, often building on work that has come before, this is immensely useful. Managing references manually adds needless complexity and room for error, and can be much better handled by these packages.

The second feature is more exotic—and some version of this functionality is utterly essential for realising DRE. *Bidirectional links* allow theorists to mimic and map RE's non-linear reasoning structure, capturing its constitutive interconnections digitally[6]. In zettelkasten applications, wrapping text in a double set of square brackets, like [[this]], converts it to an internal hyperlink the connects to another note in your collection, much like the links between Wikipedia pages. Clicking the link leads you to the other note, whilst also placing a 'backlink' which relates the target back to the origin. In effect, linking thus draws a virtual line between two notes, like string on a corkboard. Properly utilised, this functionality allows theorists to make explicit and track the connections between all the different elements of the RE process.

Let me provide an example. In the chain of notes *a* - *b* - *c*, note *b* could justify the link between *a* and *c*. In an RE referring to the trolley problem, for instance, it might read 'killing is different to letting die', and thereby connect the principle 'killing is always wrong' in note *a* with a judgement in note *c* that 'diverting the trolley is unacceptable'. Such relational associations are central to the RE method, and so links support theorists to build up webs of interrelations between judgements, principles, and theories to make their interactions more easily understandable. Importantly, this is also achieved without requiring linearity—each note can have unlimited links, allowing the ideas it contains to connect to and play a part in justifying any other idea in the coherent web as required.

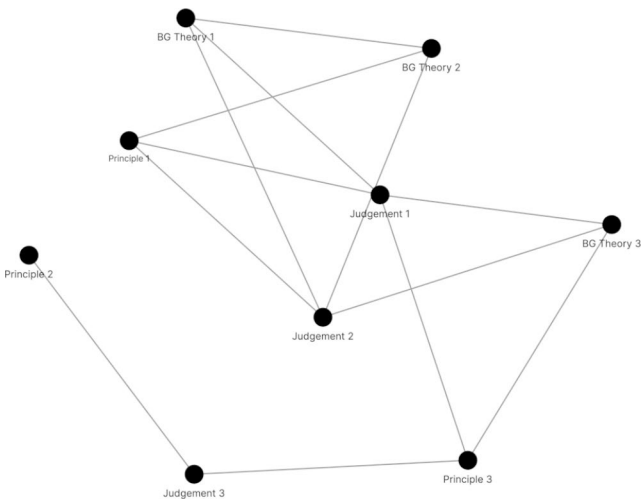

Fig. 3 A screenshot of Obsidian's graph view, demonstrating a simple web of commitments

The third feature is *tagging*. Placing a hash symbol in front of any word in these applications turns it into a #hashtag. The programs will then track these tags, so that all notes with a particular tag can be searched for, filtered and retrieved. At the very least, this would allow theorists to categorise and organise different topics in their DRE. All notes relevant to the #harmprinciple or #utilitarian approaches, for instance, could be tagged as such. If carefully managed, tags would therefore help ensure that pertinent connections between ideas with common application areas were not being overlooked. They would also allow theorists to pick out and display subsections of notes by their tags. As an RE grows, this will become an increasingly useful management aid, letting a theorist zero in on portions of their developing equilibrium without needing to consider unrelated branches.

The other way I imagine tagging being used is to identify each element as either a #judgement, #principle, or #backgroundtheory when pursuing a wide RE. I should note, however, that although my earlier sketch perhaps suggested hard and fast distinctions between the various levels of abstraction in the RE process, these distinctions are only analytical. Theorists do not necessarily need to delineate their beliefs into neat silos before embarking upon the adjustment process—"there is no problem with the boundaries between [elements] being fuzzy or overlapping" (Knight, 2017, p. 53). Indeed, part of the RE process involves refining and revising initial commitments, even moving ideas between levels of abstraction as we promote judgements to principles or vice versa. Thankfully, as notes can have multiple tags and be altered at any time, tagging flexibly supports this ambiguity.

The final feature is *graph visualisation*. Many zettelkasten programs can generate a visual representation of the theorist's web of notes, rendering any links as connections between nodes on a graph (Fig. 3). Unlike with more traditional methods of RE construction, this functionality literally allows the theorist to see, explore, and trace the connections between their ideas. Most mundanely, this can help them understand the overall 'shape' of their arguments, locating major branches and clusters. Perhaps more

[6] Unidirectional or 'directed' linking could achieve the same effect, whilst also visually encoding more information about the relationships between notes. But backlinks must then be created manually, making more work for theorists and potentially leaving room for error. Individual preference will therefore need to dictate whether theorists opt for the simplicity of bidirectional links or the more nuanced unidirectional approach.

usefully, though, theorists would also be able to more readily identify either highly connected or completely unconnected notes, which will each require attention for different reasons. Notes with many connections, for instance, will enjoy greater prima facie support from numerous neighbours, pointing to their possible nodal importance within the developing web of coherent ideas. Conversely, unconnected or orphaned ideas will need to be better linked to others, or else outright rejected due to a lack of support. Identifying both will be made far easier by digitisation, thereby aiding the construction of well-supported, coherent, and holistic arguments.

(Re)construction and explainability

Taken together, these features form the technical foundations needed for realising DRE. They will empower theorists to more thoroughly map and manipulate their ever-developing positions, with digital tools supporting greater systematicity, flexibility, and visualisability than alternatives. Much of this utility comes from being able to easily cite ideas, categorise as well as manage different elements, and link together a distributed body of notes to build up a coherent and navigable whole. By contrast, marshalling similar levels of oversight and control using word processors, whiteboards, or pens and paper would be exceedingly difficult, if not entirely unachievable. For these reasons, I suspect that the theorist working digitally is far more likely to be able to actually perform the intellectually demanding and practically extensive task of constructing an RE—particularly with regards to the substantial requirements of wide RE.

Whilst many of the principal benefits of DRE are thus essentially practical, digitisation also brings other advantages. For instance, Christoph Baumberger and Georg Brun (2021, p. 7935) have argued that one of "the most pressing internal challenges for defenders of reflective equilibrium" is the need to specify a more exact characterisation of the process of equilibrating. Although they submit that much of this will be project-dependent, Baumberger and Brun emphasise that, whatever the project, reconstructing the process of reaching an actual RE not only contributes indirectly to its justification but is also the only way for the theorist to develop "a precise configuration" of the rules for adjustment and epistemic goals that drove their theorising in the first place (Baumberger & Brun, 2021, p. 7937). In other words, it is only once a candidate RE has been found that we can retrospectively discern the most important drivers of the equilibrating process.

Working with DRE will make this evaluative task far easier. In fact, theorists that take full advantage of tagging and linking will find that much of their reconstruction work is already done, as these features can track meta-information about reasoning and decision making as they work. A theorist might, for instance, tag starting judgements as #initial-commitments, changing those tags to #discarded, #replaced, or #adjusted as and when they are supplanted (cf. Baumberger & Brun, 2021, p. 7932). Or, they might tag the #rationale for each of these decisions in a separate note, linking between these justifications and the elements they explain. Thus, in the chain of notes a_1 - b_1 - c_1, note b_1 could explain why a_1 was retired in favour of c_1. Analysing these justificatory notes in aggregate will, in turn, likely help theorists to summarise their overall epistemic goals and favoured theoretical virtues. Any notes recording such higher-level justifications will likely garner significant numbers of links and so become focal points on the graph, further attesting to their importance in the equilibrating process.

Ascertaining similar levels of insight, let alone gathering quantitative evidence akin to thickets of links, strikes me as hard to achieve with any certainty without working digitally. In this way, the explainability of an RE is enhanced by its digitisation. Digitising effectively makes all of these connections and reasons explicit, forcing reasoning out of the theorist's head and into their (digital) notes. Just as writing helps to clarify our ideas, constructing DRE will thus help to clarify the reasons for the interconnections between those ideas. This has obvious value for theorists, as it puts them in a better position to be able to explain to themselves why they made certain decisions with reference to their web of notes. However, when it comes to evaluating reasons and justifications, digitisation potentially has far more to offer to people other than the theorist. The real value of DRE comes from their potential to trigger significant reforms around research dissemination and consumption, providing a far more straightforward way for readers and critics to explore a theorist's entire position.

Disseminating equilibria

My second major claim is that we should come to expect theorists to publish entire reflective equilibria online. This supports the need to embrace DRE as standard, as the efficient distribution of equilibria can be best achieved digitally. To be clear: this is not just a case of can implies ought. Whilst the advent of the internet has rendered current research dissemination norms needlessly opaque and outdated, the increased transparency that distributing equilibria would bring should have several further advantages for both theorists and their readers. Just as the open science movement has reformed neighbouring disciplines, adopting DRE could be expected to formalise appeals to RE-based arguments in normative discourses, clearing up numerous

opportunities for misunderstanding that exist under the current distribution model.

Today, of course, entire equilibria are almost never published, either online or offline (Rechnitzer, 2022, p. 1). This is largely due to the form in which philosophical research circulates. Monographs, chapters, and papers are not only limited by their linearity, they are further confined—thanks to the costs and practicalities of putting arguments down in print—in terms of both their fixity and length. We have already seen how hard it is to convert a web of ideas premised upon mutual, coherent support into linear form. But even if publishers could find a way to print entire equilibria produced in an analogue form, not to mention come up with a viable business model for their distribution, the medium's fixity would remain an issue. Assuming equilibria should be "permanently open to revision" (Knight, 2017, p. 59), the time, money, and labour it would cost to constantly update and circulate alterations would simply be unworkable. Libraries would soon be overflowing with editions and revisions.

At the same time, we evidently cannot stick with the status quo. As it is currently not viable to publish entire equilibria, theorists are effectively forced to decide which limited fragments of an RE to highlight in print to support their arguments. Readers must then take it on trust that a coherent whole exists, somewhere behind the scenes, within which these fragmented summaries might fit. This obscures the full range of judgements, principles, and background theories that a theorist is trying to balance, as well as the argumentative moves being made to connect and develop them. In doing so, it removes all that reasoning from public view. Large portions of the most carefully-constructed equilibria will therefore never see the light of day, leaving an interested reader with little more to do than attempt to reverse engineer a theorist's RE from the few excerpts that do make it into print—essentially guessing about that theorist's wider position.

This opaqueness not only leaves plenty of room for misunderstanding otherwise well-constructed equilibria, it shields possibly weak, incomplete, or even lazy theorising from scrutiny. But, as I have already mentioned, in a discipline that ostensibly prides itself on the pursuit of argumentative clarity and rigour, a publishing model that asks interlocutors to either guess the details of arguments or blindly trust that such arguments fit into a well-constructed whole should strike us as deeply bizarre. Consequently, it seems clear to me that the lack of a workable dissemination model for reflective equilibria is creating unnecessary exegetical work for readers, forcing theorists to compensate for this limited level of disclosure in their other writings. Just think how much uncertainty could be avoided if direct references to entire DRE could be made in essays and monographs[7]! The norms and technical limitations of traditional printing practices are therefore potentially holding back contemporary RE-based discourses.

Flexibility, reproducibility, and comparison

Fortunately, a solution is already close at hand for the theorist working digitally. Most obviously, digital information is endlessly editable, meaning there is no need for the fixity of traditional publishing. Embracing DRE would therefore empower theorists to work flexibly—allowing them, for instance, to add new branches to stable and coherent positions at any time or, more ambitiously, even extensively alter existing branches and clusters of reasoning in light of new evidence. For transparency and explainability's sake this would all need to be tracked and recorded, but anything from individual principles and judgements all the way up to entire equilibria could, in theory, be readily edited or even outright replaced at will. Rather than calcifying their equilibria for print, digitally-minded theorists could thus continually revise and share works in progress, or even iterate upon published DRE after the fact, better reflecting Rawls' notion of an indefinite struggle for RE.

More importantly, digital information is also endlessly reproducible; data can be non-destructively copied and shared for next to no cost. This utterly nullifies the need for concision that printing presses once necessitated. Some of this benefit has of course already been felt in academic publishing with the advent of online access journals and publications, but it would also render the cheap and easy distribution of entire equilibria eminently possible. Indeed, I think theorists could and should be publishing DRE online—preferably via agreed upon, open formats, held in open data repositories—and explicitly referencing them in any publications. Under this new norm, rather than reconstructing a (presumably) coherent equilibrium from mere excerpts, readers would instead be able to explore the total coherence of a position, should they so wish. This would allow them to bypass all ambiguities and refer directly to the relevant RE to contextualise a theorist's arguments and understand their wider commitments[8]. Such radical transparency would

[7] A reviewer raised the possibility that pressure to regularly publish equilibria could create a kind of straightjacket effect, constraining theorists' creativity and freedom. This concern should be taken seriously and fully evaluated if the reforms to academic publishing I suggest here ever transpire. But I would stress that, in my view, theorists should only ever be publishing DRE as and when they see fit. Any theorists that felt constrained would therefore be welcome to publish at a slower cadence, keeping more of their thinking private until they were ready to share their full position later.

[8] Furthermore, if RE construction is path dependent (McPherson, 2015, p. 662), then tracing those paths with digital tools can be expected to help readers understand why certain choices were made, and where their own views might diverge.

drastically increase the clarity of RE-based discourses, bringing the benefits of externalisation, retraceability, and systematicity—not to mention graph visualisation—to the wider academic community.

Indeed, successful prior experiments with digital zettelkästen in other contexts provide reasons to think this could be eminently feasible, albeit with the caveat that they have so far all been constructed for non-RE purposes. Numerous successful projects to develop so-called 'second brains', 'digital gardens', and 'digital vaults' are currently being maintained using zettelkasten applications and uploaded to the internet for anyone to browse and critique. And many of these comprise thousands of interconnected, tagged, and referenced academic notes. These examples—which I would encourage readers to seek out if they are interested—provide compelling reasons for thinking the same should be possible in an RE context.

Why, though, would a theorist want to share their RE? And why should they care about exploring other peoples'? Quite simply, open distribution is essential for realising a vital yet hitherto neglected aspect of the RE process. To reach a wide RE, theorists should be evaluating their own position against others'—especially those with distinct experiences and outlooks—to alert them to their biases as well as any under-theorised areas of thought (Daniels, 1996, p. 2). Yet, as Baderin (2017, p. 14) notes, whilst this notion of RE comparison is "commonly put forward" in the literature, it has so far "not been adequately developed" in practice. This is surely due in part to the sheer practical impossibility of making such comparisons under the status quo. Not only are most equilibria currently never published but, even if they were, print-based distribution would be inadequate for all the reasons listed above. By contrast, a standardised system for publishing DRE would enable direct comparison, addressing one of the biggest current barriers to implementing wide RE[9].

Open scholarship

Of course, I can only suggest potential models for distributing DRE here. The wider community will need to agree on an ultimate approach. But perhaps the least inspired solution would be for the established journal publishers to host DRE in online appendices accompanying traditional publications. Sections of essays and monographs could then link through to the relevant portions of a theorist's digital RE, hugely increasing explainability. This is certainly an option. But if the whole point of wide RE is that theorists should be exposed to as many other viewpoints as possible, then to me that should be reason enough to reject locking DRE away behind the paywalls of academic publishers. Instead, the alignment of DRE with an open science ethos, which holds that research processes and outputs should be kept "transparent and accessible" and be "shared and developed through collaborative networks" (Vicente-Saez & Martinez-Fuentes, 2018, p. 434), seems a more natural fit.

What would an open approach look like? Given that DRE are essentially folders of plain-text notes, they will be relatively easy to share. Whilst theorists could therefore freely host DRE on their personal sites, I suspect many would rather focus on their theoretical work. To nonetheless facilitate an 'open humanities' or 'open scholarship' approach (Knöchelmann, 2019), a better option would be to utilise a free or not-for-profit service to manage hosting and any technical issues on theorists' behalf. Inspiration could be drawn from numerous preprint archives, or else specialist services like PhilPapers (which runs free research indexing and archive services for philosophers) or the more general GitHub (a popular open-source software host) could provide avenues for transparent and accessible DRE dissemination. Alternatively, universities and other research institutions could extend the scope of their archives and open data repositories to store DRE, too.

Whatever form distribution might take, a commitment to open access DRE can be expected to bring the benefits of the former to the latter. This would not only remove existing barriers to accessing RE-based knowledge, increasing research impact by exposing more potential interlocutors to theorists' research, but also possibly expand the range of voices that can take part in academic discussions (Willinsky, 2006, p. 32). Rather than being locked away or limited to the ivory tower, open access DRE would be widely and freely distributed over the internet. This would maximise opportunities for RE comparison, bringing with it all the chances for mutual learning and the strengthening of arguments that that would entail. This is in direct support of the aims of wide RE and, further, potentially even hints at the beginnings of a method for identifying those principles of justice that could support an overlapping consensus of the kind that emerged in the later Rawls.

This also marks a turning point in my argument. Until now, I have largely adopted the point of view of the individual theorist. Any benefits for the broader research community have been incidental—mere by-products of reconstruction or the greater flexibility and explainability of externalised equilibria. Implicitly, the information flow

[9] For similar reasons, DRE could help address issues surrounding theory acceptance. In a Rawlsian sense, theories must be able to appeal, at least in principle, to a variety of reasonable people (List & Valentini, 2016, p. 527). But how are others meant to judge whether a position is appealing? Standardised DRE publication practices would at least open the door to public understanding and the debate of various positions, even if DRE would likely require a concerted translation effort on the part of the theorist if they are to be made accessible to this wider audience.

has been one-way, from theorists to their readers, much like under the print-based dissemination model. But fully embracing open scholarship would make it entirely possible to conceive of two-way flows between theorists and their readers. It is hard not to understate the possibilities this could unlock. A digital-first approach to theorising could radically alter how theoretical work is carried out by involving others in the RE process. It is to these possible social aspects of theorising that I will now turn.

Critiquing equilibria

My third and final major claim is that normalising the open publication of DRE would allow for the auditing and critique of entire equilibria. This speaks to a different sense of reproducibility, inspired by developments in the social sciences that have sought to address the so-called 'reproducibility crisis'. The idea, briefly, that has emerged in these neighbouring fields is that a commitment to sharing research data, code, and methods allows third parties to replicate research findings and so verify results (Barba, 2018). Analogously, DRE should not only be made available to help evaluate arguments that theorists make in their other, more traditional research outputs. DRE themselves could be peer-reviewed by the wider academic community. This would allow peers to check for purported properties, like coherence, plausibility, and parsimony, and could even pave the way for possible collaborative applications of RE.

I will motivate this claim—and so illustrate the communal potential of DRE—with reference to coherence due to its centrality to RE. But various epistemic goals could be audited along the same lines. One of the stronger criticisms of RE, and coherentist accounts more generally, concerns whether or not judging coherence is practicable. The critique operates on two different levels. At the micro level, many doubt whether it is possible to judge that a particular RE is coherent, and so whether a state of RE has actually been reached. Strong (2010, p. 134), for instance, argues that RE "does not pass the test of usefulness" for this very reason. He asks (Strong, 2010, p. 133):

> [H]ow is [the theorist] to know whether she is in WRE or whether there is more adjusting to be done? […] How does she know whether she has brought into her reflective equilibrium all of the components that should be part of it? And when is it reasonable for her to believe that there are no overlooked inconsistencies?

I would add a further question: How is the theorist to know that their RE is adequately supported? Coherence, after all, is not merely concerned with consistency; all the various elements involved also need to provide each other with mutual, holistic support. Similarly, at the macro level, Ibo van de Poel (2016, p. 188) has pointed out that we need a way to operationalise comparisons between competing equilibria. If coherence is one of the major criteria by which we evaluate reflective equilibria, then we also need to be able to judge which equilibria are more or less coherent overall—a capability we currently lack. Without satisfactory ways to address these challenges, at both levels, RE is left once again looking unviable and impractical as a method.

Disappointingly, however, the standard response from RE's supporters is usually deflationary. When challenged, theorists will often emphasise RE's ideal character, admitting that whilst total internal coherence—i.e., generating an entirely coherent whole—may never be achieved in practice it is nevertheless worth striving for (Strong, 2010, p. 133). By the same token, evaluating as many competing views as possible is usually considered to be a goal worth pursuing, even if comparison with every alternative may ultimately be unachievable (Scanlon, 2002, p. 149). But is such a retreat from realising the purported central theoretical goal of RE really the only way to save it from contact with reality?

These concerns share a common root: they recognise that human beings are epistemically limited, so suggest we moderate our expectations of RE as a result. Like me, many may doubt that the lone theorist in the proverbial armchair could hold an entirely coherent RE in their heads, let alone perform the second-order task of evaluating it for coherence[10]. But I have argued that RE-based theorising should be externalised and shared precisely to ensure that equilibria are not rendered inaccessible and inscrutable to others. There is therefore no need to expect any one theorist to evaluate their particular equilibrium alone. Instead, we can appeal to a sensible division of intellectual labour to tackle the interrelated goals of judging both internal and relative coherence more feasibly. Though digitisation is no panacea, it could nonetheless enable a distributed, communal form of evaluation fit for addressing these issues.

Communal evaluation

If we accept that no theorist is an island, practically judging coherence (and other properties) becomes far more viable. An analogy can be made to existing processes of academic peer-review. With regards to the challenges from Strong and Van de Poel, imagine if theorists could upload their DRE to open access repositories to solicit communal evaluation from colleagues and peers. Third parties could then explore these equilibria and offer their opinions about whether more work is required to reach RE by submitting annotations

[10] This is only compounded by the fact that we may not even have straightforward epistemic access to our own beliefs (cf. Ryle, 1949).

and alterations. Such critical scrutiny would help theorists ascertain whether they had included all necessary elements in their RE, whether they had overlooked any inconsistencies, as well as whether each individual element was in receipt of enough support. In the spirit of wide RE, theorists would then be well-placed to go away, deliberate, and decide whether or not to incorporate any critical suggestions back into their overall position.

Such communal evaluation should be distinguished from other attempts to bring third parties into the RE process. Pioneering theorists including Miller (2003), Martine de Vries and Evert van Leeuwen (2009), as well as Jonathan Wolff and Avner de-Shalit (2007), have all sought to include others in their theorising to various extents. But such efforts have mostly involved treating public opinion as an early input to RE; a source of initial judgments and/or principles. Whilst some have gone further, proposing to involve multiple people in RE's balancing and revisioning processes (cf. Savulescu et al., 2021, pp. 657–659), I agree with Baderin (2017, p. 12) that this seems unworkable in practice. Garnering public input to an RE may well be desirable and achievable, but notions of large-scale collective theorising seem to underappreciate just how difficult—if not impossible—it would be to actually reconcile nuanced, often-conflicting worldviews into a mutually acceptable and shared whole.

To clarify, I am not suggesting that theorists should collaborate with their peers to reach collective DRE. Peer-reviewers would not strive for a shared consensus, and nor would they need to internalise any equilibria they evaluated. They would not even need to agree with the theorist's choice of inputs, though should probably propose any overlooked ideas for consideration. Instead, my proposal here is best understood as an attempt to split the task of RE evaluation up into more manageable chunks. This is intended to help deal with the complexity and scale of DRE. At the micro level, for instance, justice theorists could draw on their expertise to scrutinise any justice-related elements of a position, whilst metaethicists could focus on areas relevant to their own proficiencies. At the macro level, we could then aggregate all these judgements to help ascertain which equilibria were more or less coherent overall. Though care would evidently need to be taken to avoid reductionism, I nonetheless contend that this communal approach offers a potentially workable method for evaluating coherence. Determining the consistency of a local cluster of related ideas (and their various interconnections) seems far more viable for individuals than judging global coherence.

Naturally, however, any chorus of critical voices is unlikely to be entirely harmonious. Peer-review will thus provide the theorist with many considerations to reflect upon, some of which may even conflict. But this is not an issue. As they retain ultimate control over their position, the theorist will simply need to decide what to accept and integrate into their RE, or else provide reasons (perhaps publicly) for rejecting critics' suggestions. Much like with the manuscript peer-review process, such efforts would amount to a form of loosely democratic quality-control (Knöchelmann, 2019, p. 9). Compared to the status quo, where there is limited transparency around reflective equilibria and almost no scope for their critique, the potential gains should be clear. Even if all errors, inconsistencies, and insufficiently supported elements are not uncovered, exposing DRE to far more critical scrutiny can be expected to increase their quality.

Collaboration, some concerns and suggestions

That said, there is no need to completely foreclose all collaborative possibilities. Whilst I personally doubt how effective collective theorising could be, these proposals do provide scope for adventurous theorists to explore aspects of cooperative working—a by-product of the greater accessibility and flexibility of DRE. Under the open access model, two or more theorists could plausibly come together and work towards a shared RE, much like multiple authors collaborating on a Google Doc. As mentioned earlier, though, the likelihood of two theorists completely agreeing on all the necessary judgements and principles, not to mention mutually accepting all revisions, strikes me as vanishingly improbable. But such prospects could nevertheless pose an interesting avenue for future research to explore.

Perhaps a more realistic possibility is that third parties could branch off from a theorist's RE to create derivative positions where there is divergence in only some aspects of thought. Each theorist would then still be responsible for managing their own RE. A student might largely agree with their professor, for instance, but wish to change one important principle. After duplicating their professor's position, they would then need to refactor any ideas which the new principle touched upon. This would likely require significant knock-on alterations to render the RE coherent again. Along similar lines, when encountering compelling alternative views we might imagine merging portions of a theorist's position into our own[11]. Lacking expertise in jurisprudence, I could defer to a more qualified theorist on matters relating to the law, incorporating their insights into my own RE. In this way, working digitally could potentially enable us to 'remix' new equilibria.

However, feasibility concerns aside, I am reticent to fully endorse these more radical proposals. An undesirable corollary of publishing DRE in this level of detail is that it will also make it easier to plagiarise views. Academics

[11] This could be one way to operationalise the process of reaching and establishing an overlapping consensus, as outlined in Rawls' *Political Liberalism* (2005b).

have, over centuries, developed robust norms for acknowledging the ownership of ideas, but we are already seeing these practices strain in the globalised knowledge economy. Any developments, like DRE, that might make it easier for unscrupulous individuals to pass ideas off as their own thus need to cautiously implemented. Even where there is no malicious intent, like with my jurisprudence example, mechanisms for tracking intellectual property would need to be properly engineered. Solutions would thus require the cooperation of zettelkasten software developers, who would need to build ways of tracking authorship metadata into their applications. This may not be an insurmountable issue, but will need to be tackled if DRE are to proliferate.

If software developers are interested in altering their applications to better support DRE, there is one other modification I would also implore them to make. Tracking the development of ideas over time is currently a challenge in existing tools. I have suggested that, as notes are created, edited, and replaced, these changes and their justifications need to be recorded. Without such evidence, it is hard to evaluate how a theorist's position has evolved and developed. Something akin to the 'version control' systems that track code revisions in software development would thus be useful here. Embedding similar systems within programs designed for constructing DRE could help ensure that any changes, and the reasons for making them, can be understood by third parties once an RE is published. It would also help theorists to manage their DRE as they develop, affording them yet more ways to clarify their thinking.

Conclusion

Analytic (political) philosophers claim to be concerned, above almost all else, with the pursuit of argumentative rigour and clarity. At the same time, many claim RE is their primary methodological tool. Yet, under the status quo, the full details of theorists' equilibria—assuming they exist at all—are left cloaked in darkness, directly obstructing those pursuits. This paper has accordingly provided pragmatic guidance for the theorist looking to bring their RE into the light. Motivated by a perceived need for greater transparency and accessibility, as well as a better way to tackle the ambitious task of actually constructing thoroughgoing equilibria, I have explained how to repurpose off-the-shelf tools to construct DRE and defended the practical and theoretical benefits of working digitally. These programs already exist, and are waiting for trailblazing theorists to deliver on the method's full promise.

The next challenge will be to construct a digital RE. Whilst it is easy to describe the process in theory, it will require significant work in practice—as well as access to a standardised system for distributing equilibria along the lines outlined above if a wide RE is likely to be achieved. As this demonstrates, no theorist can deliver on the full promise of RE alone. It is only by combining RE with digital construction techniques, networked communication technologies, and new publication norms that we can expect to fully realise the method's latent possibilities. Though the thorough construction of DRE (especially of the wide variety) is still sure to be time-consuming and intellectually demanding, even with the help of digital tools, I hope to have nevertheless demonstrated how digitising the process could feasibly unlock real methodological value for interested theorists.

Acknowledgements I would like to thank Lesley Smith and Laurel Boxall, along with Georg Brun, Tanja Rechnitzer, Gregor Betz, Wibren van der Burg, and all the other attendees of 'Reflective Equilibrium: *51 Years after A Theory of Justice*' for their thought-provoking feedback on an earlier draft of this paper.

Funding This work was supported by the Economic and Social Research Council [Grant Number: ES/P000649/1].

Declarations

Competing interests The author has no financial or proprietary interests in any material discussed in this article.

References

Baderin, A. (2017). Reflective equilibrium: Individual or public? *Social Theory and Practice*, *43*(1), 1–28. https://doi.org/f3s8gx.

Barba, L. A. (2018, February 9). *Terminologies for Reproducible Research*. Retrieved from http://arxiv.org/abs/1802.03311.

Baumberger, C., & Brun, G. (2021). Reflective equilibrium and understanding. *Synthese*, *198*, 7923–7947. https://doi.org/ggkp4w.

Cath, Y. (2016). Reflective Equilibrium. In H. Cappelen, T. S. Gendler, & J. Hawthorne (Eds.), *The Oxford Handbook of Philosophical Methodology* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199668779.013.32.

Daniels, N. (1979). Wide reflective equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy*, *76*(5), 256–282. https://doi.org/10.2307/2025881

Daniels, N. (1996). *Justice and justification: Reflective equilibrium in theory and practice*. Cambridge University Press.

De Vries, M., & Van Leeuwen, E. (2009). Reflective equilibrium and empirical data: Third person Moral Experiences in empirical Medical Ethics. *Bioethics*, *24*(9), 490–498. https://doi.org/10.1111/j.1467-8519.2009.01721.x.

Doorn, N., & Taebi, B. (2018). Rawls's wide reflective equilibrium as a method for engaged interdisciplinary collaboration: Potentials and Limitations for the Context of Technological Risks. *Science Technology & Human Values*, *43*(3), 487–517. https://doi.org/10.1177/0162243917723153.

Elgin, C. Z. (1999). *Considered Judgment*. Retrieved from https://press.princeton.edu/books/paperback/9780691005232/considered-judgment.

Erz, H. (2023). Zettlr. Retrieved from https://doi.org/10.5281/zenodo.2580173.

Erz, H. (2019, October 26). Release: Developing Open Source is a Political Act. Retrieved March 6, 2022, from https://www.zettlr.com/post/release-developing-open-source-political-act.

Knight, C. (2017). Reflective equilibrium. In A. Blau (Ed.), *Methods in Analytical Political Theory* (pp. 46–64). Cambridge University Press.

Knöchelmann, M. (2019). Open Science in the Humanities, or. *Open Humanities? Publications*, *7*(4), 65. https://doi.org/10.3390/publications7040065.

Lewis, D. K. (1983). *Philosophical Papers* (1 vol.). Oxford University Press.

Li, S., & Xu, E. (2023). Obsidian. Retrieved from https://obsidian.md/.

List, C., & Valentini, L. (2016). The Methodology of Political Theory. In H. Cappelen, T. S. Gendler, & J. Hawthorne (Eds.), *The Oxford Handbook of Philosophical Methodology* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199668779.013.10.

McPherson, T. (2015). The methodological irrelevance of reflective equilibrium. *The Palgrave Handbook of Philosophical Methods*, 652–674.

Miller, D. (2003). *Principles of social justice* (3. print.). Harvard University Press.

Numerous Authors (2023). Logseq. Retrieved from https://github.com/logseq/logseq.

Rawls, J. (1951). Outline of a Decision Procedure for Ethics. *The Philosophical Review*, *60*(2), 177–197. https://doi.org/10.2307/2181696.

Rawls, J. (1974). The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association*, *48*, 5–22. https://doi.org/10.2307/3129858.

Rawls, J. (2005a). *A theory of Justice*. Harvard University Press. (Original Edition).

Rawls, J. (2005b). *Political liberalism*. Columbia University Press.

Rechnitzer, T. (2022). *Applying reflective equilibrium*. Springer.

Ryle, G. (1949). *The concept of mind*. University of Chicago Press.

Savulescu, J., Gyngell, C., & Kahane, G. (2021). Collective reflective equilibrium in practice (CREP) and controversial novel technologies. *Bioethics*, *35*(7), 652–663. https://doi.org/10.1111/bioe.12869.

Scanlon, T. M. (2002). Rawls on Justification. In S. Freeman (Ed.), *The Cambridge Companion to Rawls* (pp. 139–167). Cambridge University Press. https://doi.org/10.1017/CCOL0521651670.004.

Strong, C. (2010). Theoretical and practical problems with wide reflective equilibrium in bioethics. *Theoretical Medicine and Bioethics*, *31*(2), 123–140. https://doi.org/dvn6qz.

Van de Poel, I. (2016). A Coherentist View on the relation between Social Acceptance and Moral Acceptability of Technology. In M. Franssen, P. E. Vermaas, P. Kroes, & A. W. M. Meijers (Eds.), *Philosophy of technology after the empirical turn* (23 vol., pp. 177–193). Springer International Publishing. https://doi.org/10.1007/978-3-319-33717-3_11.

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043.

Willinsky, J. (2006). *The access principle: The case for open access to research and scholarship*. MIT Press.

Wolff, J., & de-Shalit, A. (2007). *Disadvantage*. Oxford University Press.