Can AI Rely on the Systematicity of Truth? The Challenge of Modelling Normative Domains

MATTHIEU QUELOZ

Abstract: A key assumption fuelling optimism about the progress of Large Language Models (LLMs) in modelling the world is that the truth is *systematic*: true statements about the world form a whole that is not just *consistent*, in that it contains no contradictions, but *cohesive*, in that the truths are inferentially interlinked. This holds out the prospect that LLMs might rely on that systematicity to fill in gaps and correct inaccuracies in the training data: consistency and cohesiveness promise to facilitate progress towards *comprehensiveness* in an LLM's representation of the world. However, philosophers have identified reasons to doubt that the truth is systematic across all domains of thought, arguing that in normative domains, in particular, the truth is not necessarily systematic. I argue that insofar as the truth in normative domains is asystematic, this renders it correspondingly harder for LLMs to make progress, because they cannot rely on the consistency and cohesiveness of the truth to work towards comprehensiveness. And the less LLMs can rely on the systematicity of truth, the less we can rely on them to do our practical deliberation for us, as there is correspondingly more of a role for human agency in navigating asystematic normative domains.

Keywords: AI; language models; normativity; value pluralism; conflicts of values; agency; theory of action; hard choices; consistency; coherence; authenticity.

Word count: 7500 words excl. notes; 8100 words incl. notes.

1. Introduction

Asked how Large Language Models (LLMs) could ever hope to comprehensively model the world as long as they were trained on incomplete and not fully accurate data, Dario Amodei, co-founder and CEO of Anthropic, invoked a time-hallowed idea as a reason for optimism, namely that *truths are integrated in a systematic web*:

> There's some relatively simple web of truth out there in the world … all the true things are connected in the world, whereas lies are kind of disconnected and don't fit into the web of everything else that's true. (Amodei 2024)

We should not expect incompleteness and inaccuracies in the training data to be an insuperable obstacle to progress in AI, Amodei is suggesting, because LLMs can work their way from those partially inconsistent data to the simpler, unified web of truth that underlies them, and rely on the systematicity of that web of truth to fill in lacunae and smooth out noise or inaccuracies in the training data. The web of truth promises to allow LLMs to jump beyond the data they are trained on while acting as a safety net preserving them from error.

The crucial assumption fuelling this optimism is that the truth is *systematic*: the totality of true statements about the world forms a whole that is not just *consistent*, in that it contains no contradictions, but *cohesive*, in that the truths it consists of are inferentially interlinked: they stand to each other in relations of material implication and exclusion that allow a given truth within the system to be inferred from other truths in the system. This cohesiveness offers grounds for a certain optimism because it holds out the prospect that truths that are missing or misrepresented in those data might be derived from the truths already represented in the model. In other words, the consistency and cohesiveness of the truth together facilitate progress towards *comprehensiveness* in LLMs' modelling of the world.

Yet philosophers have identified various reasons to doubt that the truth is systematic across all domains of thought. In particular, value pluralists have offered forceful arguments to the effect that *in normative domains*, the truth is *not* necessarily systematic: statements expressing values, or describing how the world *ought to be* rather than how it is, do not necessarily fit together into a consistent, cohesive whole. And insofar as the truth in normative domains is not systematic, this renders it correspondingly harder for LLMs to make progress in these domains, because they cannot then rely on the consistency and cohesiveness of the truth to work towards comprehensiveness.

In a nutshell, my argument is therefore the following: Where truth is systematic, progress towards comprehensiveness is in principle facilitated by the fact that LLMs can rely on the systematicity of truth to extrapolate from incomplete training data. But where truth is not systematic, progress is likely to be hampered by the fact that LLMs cannot rely on the systematicity of truth to extrapolate from incomplete training data to the same extent. In normative domains, there are reasons to think that the truth is not systematic. Therefore, there are reasons to think that LLMs will find it significantly harder to comprehensively model truths in normative domains than truths displaying systematic integrity. The less LLMs can rely on the systematicity of truth, moreover, the less we can rely on them to do our practical deliberation for us, because there is correspondingly more of a role for human agency to play in navigating asystematic normative domains.

2. The Systematicity of Truth

The idea that truths interlock to form a systematically integrated whole has a long history in philosophy. The Stoics already referred to the *systema mundi*, the system of the world, and modern philosophers from Leibniz through Wolff, Lambert, and Kant to Hegel and

Whitehead have elaborated this idea in various ways.[1] But what exactly does it mean to say that truths form a systematic web? The basic idea is that truths are not disconnected fragments. They interlock to form a unified whole—what the original compiler of the *Oxford English Dictionary*, James Murray, called "the fabric of fact".

This means, first of all, that truths are *consistent* with each other: they are free of contradictions. This includes both direct contradictions ($P \wedge \neg P$) and indirect contradictions, where the implications of a statement turn out to contradict each other ($(P \rightarrow Q) \wedge (P \rightarrow \neg Q)$).

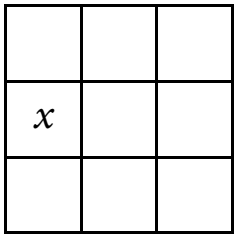
But consistency is a relatively weak requirement. A collection of disparate and unrelated truths might be free of contradictions without being systematically integrated. Moreover, consistency is easily achievable through the addition of further premises or qualifications (think of how, in the Ptolemaic model of the universe, the addition of ever more epicycles served to patch up observed inconsistencies).

When truths are systematically integrated, they are not merely consistent, but also *cohesive*: they stand in relations of rational support to one another. This inferential interconnectedness reflects an underlying interconnectedness in the fabric of facts: the obtaining of one fact implies the obtaining of certain other facts, and excludes the obtaining of yet other facts. If New York is east of San Francisco, then this implies that

[1] The richest historical overview of the ideal of systematicity in philosophy is Otto Ritschl's *System und systematische Methode in der Geschichte des wissenschaftlichen Sprachgebrauchs und der philosophischen Methodologie* (1906). It is complemented by Messer (1907), and Rescher (1979, 3–8; 2005, 19–38) offers a lucid Anglophone account of the resulting picture. Losano (1968) and Troje (1969) also examine the development of the demand for systematicity, but in the context of law and jurisprudence. Stein (1968) sketches a brief history of the concept of system, and there are also various historical contextualizations of the notions of system at work in the thought of individual philosophers: Rescher (1981) examines the concept of system in Leibniz's work; Vieillard-Baron (1975) traces the concept of system from Leibniz to Condillac; Kambartel (1969) reconstructs the notions of system and justification in Kant's work, as do Rescher (2000, 64–98), Kitcher (1986), Guyer (2003, 2005), Abela (2006), and Ypi (2021); on Hegel and systematicity, see the essays in Brooks and Stein (2017), especially Thompson (2017). Franks (2005) explores the role of the demand for systematicity in German idealism.

San Francisco is west of New York, and it excludes that New York is west of San Francisco. Accordingly, truths about geography have to be rationally coordinated with each other. Acknowledging one truth about geography provides reasons for recognizing or ruling out other truths about geography.

As Nicholas Rescher (2005) has argued, this inferential interconnectedness introduces a useful form of *redundancy* into one's understanding of the world: given a sufficient number of truths about something, any one truth could be excised from the list and still remain derivable from the rest. Rescher (2005, 5) offers a perspicuous illustration using a simple tic-tac-toe-like situation:



The following truths can be formulated about this situation, Rescher notes:

(1) There is exactly one *x* in the configuration.

(2) This *x* is not in the first row.

(3) This *x* is not in the third row.

(4) This *x* is not in the second column.

(5) This *x* is not in the third column.

(6) This *x* is not on a diagonal.

(7) This *x* is not at column-row position (3, 2).

Were one to excise one truth from this list—say, (5), "This *x* is not in the third column"—it could still be recovered from what remained. Excising (5) would still leave one able to infer it from (1), (2), (3) and (7).

As long as the fabric of fact is logically unified in this manner, a representation of that fabric of fact will thus contain redundancies that render it more *robust in the face of*

information gaps: any truth that is lost, or perhaps even missing to begin with, can nonetheless be recovered from other truths about the fabric of fact.

This cohesiveness of systematically integrated truths is relevant to the prospects for training LLMs, because it supports the *expansion of understanding by inference*. In particular, it promises to greatly facilitate *filling in what is missing* from the training data. Whether current LLMs can already reason their way to truths that are not explicitly present in their training data is contested. But the point is that the systematicity of truth supports this possibility *in principle*. Thanks to the systematic integration of the fabric of fact, truths that are not represented in the training data can in principle be recovered by inference. The systematicity of truth facilitates self-completion.

Up to a point, the cohesiveness of systematically integrated truths also promises to help LLMs in *correcting inaccuracies* in the training data. This is what Amodei is driving at when he remarks that "lies are kind of disconnected and don't fit into the web of everything else that's true." If truths about a given domain are systematically integrated, then a candidate truth's *integrability with* what is already taken to be true can act as a *criterion* for whether the candidate truth should be accepted.

Of course, lies—or, more broadly, untruths—may, though disconnected from the web of what is in fact true, nonetheless be connected to other untruths. As any good liar knows, the hardest part of lying well is to preserve the appearance of systematic integrity across a web of lies. For, if the fabric of fact is systematically integrated, then the acceptance of an untruth will require corresponding changes in other parts of one's web of beliefs, which will require further adjustments in turn, and so on, so that even one false belief is capable of overturning many true beliefs, propagating falsity through the web.

Any supposition contrary to fact thus threatens to pervasively vitiate one's understanding of the world. This has been called *Burley's principle* (Rescher 2005, 4), after the medieval philosopher Walter Burley, who observed that whenever a false contingent proposition is posited, any false proposition that is compatible with it can be derived from

it (Kretzmann and Stump 1989, 391). Given the acknowledgement of any two non-equivalent truths P and Q, we can derive from them the truth of ¬(¬P ∧ Q), which is logically equivalent to P ∨ ¬Q. But if we now assume only one untruth, say, ¬P, then there is no stopping there, because this at once implies ¬Q. Hence, it would seem that the acceptance of any one untruth (¬P) has the consequence that any other arbitrary truth (Q) has to be abandoned.

Burley's principle also has implications for LLMs. It highlights that besides the question of whether LLMs can rely on the systematicity of truth in principle, there is also the question of whether they succeed in relying on it in practice: do LLMs in fact manage to cotton on to the fabric of fact? Burley's principle underscores the real danger of their weaving together a web of falsehoods as a result of extrapolating even from as little as a single untruth. This reminds us of how easy it is for the web of truths to become obscured by a tissue of lies.

However, the fact that the assumption of a single untruth has systematic ramifications that are bound eventually to come into conflict with facts one knows to obtain can also be turned into an advantage in rendering LLMs truthful: it can be used to determine *which* systematically integrated web one should accept. Consider a detective trying to reconstruct a crime based on the testimony of various witnesses. Even if several of the witnesses are complicit and collude to entangle the detective in a systematic web of lies, that web of lies is likely to end up contradicting something that the detective, after a thorough investigation, comes to regard as incontrovertibly true. Even if the detective initially takes the misleading testimony at face value, therefore, the demand for all the facts of the case to fit together into a unified whole leads the web of lies to unravel in the end, because the systematicity of truth gives the detective *critical leverage* over the epistemic authority of the witnesses: it enables the detective to retroactively re-evaluate the trustworthiness of witnesses in light of how well their testimony *fits in* with what gradually emerges as the most consistent and cohesive account of what happened.

Analogously, LLMs could *in principle* rely on the systematicity of truth to retroactively re-evaluate inaccuracies in their training data in light of their overall fit with the most consistent and cohesive web of statements about the world. And when competing webs are equally consistent and cohesive, components of the training data that are given special weight as certifiably authoritative sources can be treated as ground truths and act as tie-breakers. Thanks to the systematicity of truth, moreover, these high-quality inputs can be used not merely to overturn statements that directly contradict them, but can be leveraged more widely. When systematically integrated, their implications can reveal tensions with the implications even of seemingly unrelated statements. The systematicity of truth, especially when supported by such a reliable basis, thus facilitates self-correction.

The upshot is that the systematicity of truth promises to be a significant aid to the development of remotely comprehensive LLMs, because that systematicity is something that LLMs can in principle rely on for *self-completion* and *self-correction*.

As we shall see in the next section, however, philosophers have identified various reasons to doubt that the truth is systematic across all domains of thought.[2] *Normative* domains, especially, which notably encompass ethics and politics, represent a complex landscape of often conflicting values, ideals, virtues, and principles. This has implications for the prospects of LLM self-completion and self-correction in these domains.

3. Value Pluralism and the Asystematicity of Normative Domains

The tradition of *value pluralism*, whose acute relevance to AI was recently underlined by John Tasioulas (2022), offers forceful arguments to the effect that when it comes to truths about what is valuable, the truth may well be asystematic: statements expressing values,

[2] Cummins et al. have also made the congenial observation that while "some domains are cognized via a grasp of their underlying structure," "some domains are cognized without grasping any significant underlying structure" (2001, 174–175). However, their conception of domains and what it takes for them to be structured differs starkly from the one at issue in what follows.

or describing how the world *ought to be* rather than how it is, do not necessarily fit together into a consistent, cohesive whole.

Value pluralism is best understood as contrasting with value *monism*. For our purposes, value monism can be thought of as involving four claims. First, there is ultimately only *one* thing that is intrinsically valuable—hence 'monism'. Second, because there is ultimately only one thing that is intrinsically valuable, everything that is good or valuable can be integrated into a harmonious whole. There are no truly irreducible conflicts between values, no trade-offs that cannot be resolved without loss, no ineluctable moral dilemmas. This is a commitment to the *compatibility* of values. Third, everything that is good or valuable is commensurable, because all aspects of what is good or valuable can ultimately be converted into a common underlying currency of value, a single metric in terms of which they can be measured and compared. This is a commitment to the *commensurability* of values. Given these three commitments, a fourth one naturally follows, namely that all *truths about* what is good or valuable are themselves compatible and commensurable, so that they interlock to form a systematically integrated whole. This is a commitment to the *systematicity of truth* about values. A paradigmatic form of value monism is utilitarianism, which takes the overarching master value in terms of which everything else is ultimately measured to be some form of *well-being*, such as preference satisfaction.[3] The influential computer scientist Stuart Russell (2019) advocates such a preference-based utilitarianism as the basis for building AI models, for example. Similarly, the psychologist and neuroscientist Joshua Greene has advocated utilitarianism on the grounds that it provides a common currency in terms of which to construct "a unified system for weighing values" (2013, 15). Such a unified system should form the basis for AI models' understanding of the normative landscape, Greene believes: "Before we put our

[3] At least, this applies to the best-known elaborations of utilitarianism. As Sen (1981) argues, it is in principle possible to envisage a value pluralist elaboration of utilitarianism.

values into machines," he writes, "we have to figure out how to make our values clear and consistent" (2016, 1515).

By contrast, value pluralism rejects all four claims characteristic of value monism. First, pluralists hold that there is not just one, but a *plurality* of irreducibly distinct values. Second, pluralists maintain that these different values are in many cases incompatible. As Bernard Williams articulates the view, pluralists consider it a "deep error" to suppose "that all goods, all virtues, all ideals are compatible, and that what is desirable can ultimately be united into a harmonious whole without loss" (Williams 2013, xxxv). Our values are bound to end up pulling in competing directions when pursued in concert, not merely because time is short or the world recalcitrant, but because the values themselves inherently conflict.[4] We face inevitable trade-offs between diverging ends, the realisation of some of which can only be obtained at the expense of others (Berlin 2002, 213–214). This is not just the Rawlsian claim that the values of some groups in society clash with the values of other groups. It is the stronger claim that even the values of a single individual clash in ways that are not resolvable without loss: think of the notorious tension between the ancient virtue of excellence and the Christian virtue of humility; or of the tensions between truthfulness and kindness, liberty and equality, loyalty and honesty, tradition and progress, justice and mercy, or security and privacy.

Third, pluralists hold that these different values are in many cases *incommensurable*. This means that when values conflict, there is no common currency in terms of which to compute the gains and losses involved in trading one value against another. But it also means something wider, namely that 'there is no other determinate and general procedure for solving conflicts, such as a lexical priority rule' (Berlin and Williams 1994, 306). Resisting the pressure to come up with techniques for making incommensurable values

[4] See Berlin (2013, 12) as well as Berlin and Williams (1994). Further elaborations of the pluralist outlook include Larmore (1987), Stocker (1990), Kekes (1993), Chang (1997a, 2015a), and Dancy (2004). See also Chang (2015b), Heathwood (2015), Mason (2023), and Blum (2023) for overviews.

commensurable, pluralists hold that our efforts 'should rather be devoted to learning—or learning again, perhaps—how to think intelligently about conflicts of values which are incommensurable' (Williams 2001, 89). One important step in that direction is to realize that while commensurability implies comparability, *incommensurability* does *not* imply *incomparability*. 'Comparison', Ruth Chang emphasizes, 'does not require any single scale of units of value' (1997b).[5] Moreover, the lack of a single scale of value does not render arbitrary an agent's judgements concerning which of two incommensurable values is more important in a given connection. Judgements of importance need not be any less rational or reasonable simply because they do not rely on a common currency of value. One can still have *reasons* to think that one value should prevail over the other in a given situation—it is merely that these reasons will not take the form of a common currency or a lexical priority rule, and will be more context-sensitive than these highly general procedures for solving value conflicts.

Fourth and finally, pluralists hold that we should not necessarily expect truths about values to interlock in a systematically integrated whole. As Thomas Nagel observes, 'truth in science, in mathematics, or in history has to fit together in a consistent system', but 'our evaluative beliefs are not part of the attempt to describe a single world' (2001, 108–9).[6] If many of our different values are genuinely distinct, incompatible, and incommensurable, the relationship between *truths about* these values becomes correspondingly complex and conflictual. It becomes possible for there to be *conflicting truths* about a given situation.

These conflicting truths can take one of two forms.[7] In the first case, one and the same action appears to be one I *ought* to perform in view of some of its features and at the same time appears to be one I ought *not* to perform in view of some of its other features: in light

[5] Chang (2015b, 25) maintains that even monism does not strictly entail comparability, because different *qualities* of a single value need not be comparable; indeed, on her account, even different *quantities* of a single value need not be comparable.

[6] See also Chappell (2009) and Hämäläinen (2009, 548).

[7] I draw here on Williams's (1973, 171) discussion of conflicts of ought.

of value x, I ought to ϕ; but in light of value y, I ought not to ϕ. Deciding whether or not to ϕ then requires one to judge the relative importance of the features that count for and against the action in this situation.

In the second case, there are two actions I each ought to perform, but I cannot perform both: in light of value x, I ought to ϕ; but in light of value y, I ought to ψ instead, and I cannot do both. The fact that I cannot do both may be due to a contingent empirical feature of the world—jackhammers working as they do at present, it may be impossible to be a jackhammer operator while moonlighting as a concert pianist.[8] But equally, the conflict may be inherent to the values themselves. There may be an inherent tension between x and y—liberty and equality, say, or truthfulness and happiness, or security and privacy—in that the sustained realization of x can only happen at the expense of the sustained realization of y, and vice versa. Again, a judgement of importance seems to be required to decide which value should prevail in a given situation, and how far one should go in sacrificing the realization of the other value to that end.

Pluralists maintain that such conflicting truths cannot always be analysed away; the conflict between them may be genuine and irreducible. This highlights a stark asymmetry between systematic and asystematic truth. When one discovers a conflict between two beliefs about a systematic domain, the discovery of the conflict is normally taken as evidence of some *epistemic error*, and one's confidence in the conflicting beliefs is *weakened* until the error is found and at least one of the two offending beliefs is abandoned altogether. But in asystematic domains, the discovery of conflict need not be evidence of epistemic error, and one's confidence in the conflicting beliefs need not be weakened at all. Rather, one's discovery of the conflict makes one realize that one faces a dilemma, or at least a trade-off, and that this requires a judgement about what is more important in that particular situation. And once that judgment has been made and one belief has

[8] I take the example from Millgram and Thagard (1996, 73).

prevailed over the other, the overruled belief does not disappear. Instead, it now registers as *regret at the real costs incurred* in terms of the value that was overruled in the name of some other value that struck one as more important in that situation.[9] Both conflicting beliefs thus endure—though after one prevails, the other resurfaces in a different guise: as regret, or as a sense of loss, which acknowledges the reality and force of the consideration that was not acted upon, and which may subsequently motivate further action, such as showing remorse, making amends, issuing an apology, or offering some sort of compensation or reparation.[10]

These ineliminable conflicts between truths about values imply that, at the end of the line, there may be no systematic harmony to be had in normative domains: the various normative truths expressing our values and describing what the world ought to be like do not fit neatly into a unified, consistent, and cohesive system. Truths about values, we might say, are at least partly *asystematic*. We constantly acknowledge this asystematicity when, by showing regret or remorse, we acknowledge that a real loss was incurred as a result of our doing something we nevertheless *had* to do. If all truths about normative domains fit together into a harmonious whole, we could realize all our values without painful trade-offs or remainders. To acknowledge that we cannot do so is to acknowledge that these truths are to some extent asystematic.

The point can be vividly put in terms of the distinction between a map and the landscape it depicts. Truths in certain domains, such as geography, form a systematically integrated whole, because the landscape they map itself forms a systematically integrated whole—in this case, the Earth, whose *ontological* systematicity ensures the *logical* systematicity of geographical truths such as "Paris is west of Prague" and "Prague is east of Paris." But if pluralists are correct, then truths in other domains, such as normative

[9] These features of value conflicts are discussed in detail in Williams (1973).

[10] On regret or a sense of loss as an acknowledgement of genuine conflicts of values, see Williams (1973, 1981a, b, 2005a), Queloz (2024), and Cueni (2024).

truths about ethics and politics, do not form a systematically integrated whole, because the landscape they map itself does not form a systematically integrated whole; rather, it forms a fragmented, tension-ridden, disparate, and disconnected landscape. This should come as no surprise if we regard the normative landscape as the historical deposit of a variety of influences and vastly different traditions.[11] Why should we expect the vicissitudes of cultural history, with all the disparate traditions of normative reflection they jumbled together and continually reconfigured in often contingent and messy ways, to produce, of all things, a practice of normative reflection tracking a neatly integrated normative landscape? Such a process is far more likely to have produced a disconnected patchwork of conflicting normative considerations. Those who insist that the normative landscape itself already exhibits perfect systematicity, pluralists maintain, owe us an explanation of how that systematicity is supposed to have got there.[12]

Insofar as truths about values are asystematic, the proper orientation towards them is not to shoehorn them into a unified and coherent system, but to aim at a nuanced understanding of the subtle interplay and trade-offs between them. As Nora Hämäläinen emphasizes, striving for systematicity and coherence may not be the "proper orientation in the moral realm," because "the gaps and leaps in our moral vocabularies and frameworks may be essential to the object of investigation—morality—rather than faults in our understanding of it, that need to be corrected by a more coherent, unitary perspective" (2009, 548). Our evaluative beliefs are accordingly free to be as irreducibly disparate, inconsistent, and tension-ridden as our values themselves are—indeed, our evaluative beliefs *should* mirror this asystematicity if they are to be true to our values.

The asystematicity of normative truths in turn has implications for the prospects of LLMs. LLMs excel at identifying and leveraging patterns and are becoming increasingly good at making inferences within systematic domains. But any machine learning

[11] See MacIntyre (2007) and Williams (2005b, 136–37).

[12] See Williams (1995c, 189).

approach that relies on identifying and leveraging systematic and consistent patterns will have a harder time modelling a normative landscape that lacks systematicity and consistency. If pluralists are right and normative domains are at least partly asystematic, then attempts to model human values cannot expect to receive the kind of support from the systematicity of truth that they can in principle rely on in systematic domains such as geography. If normative truths are asystematic, these truths will not exhibit the same degree of inferential interconnectedness and redundancy that geographical truths exhibit. When confronted with the inherent and irresoluble conflicts that value pluralism highlights, therefore, LLMs' ability to extrapolate from incomplete training data and comprehensively model human values will be hampered. If pluralists are right, the asystematicity of truth in normative domains is a significant hurdle for AI models.

This is a different hurdle from the one that Sorensen et al. (2024) have recently identified. Their worry is that insofar as AI systems are statistical learners that aggregate vast amounts of data and fit it to averages, they are ill-poised to learn about inherently conflicting values, because they risk "washing out" (Sorensen et al. 2024, 19937) just the value conflicts we want them to model. The same "washing out" problem also afflicts the attempt of Feng et al. (2024) to extract normative requirements from their LLM's pre-training data: they filter the normative requirements to ensure their consistency. But if the normative truths that LLMs are supposed to model are themselves inconsistent, this filtering process effectively distorts the model's grasp of the normative landscape it is trying to map. When dealing with asystematic domains, the very strategy that promises to help LLMs self-complete and self-correct in mapping out systematically integrated domains thus turns into a counterproductive strategy that risks distorting the map.

As Sorensen et al. (2024) show, a significant step towards overcoming this difficulty is to accommodate value conflicts by explicitly representing them within a dataset such as ValuePrism. This expressly "value pluralist" dataset leverages GPT-4's open-text generative capabilities to make explicit the wide variety of human values that are encoded

in its pretraining data. The resulting dataset covers 218k values, rights, and duties contextualized in terms of 31k human-described situations (obtained by filtering 1.3M human-written situations sourced from the Allen Institute's Delphi demo).

Trained on such a data set, models such as Value Kaleidoskope (Kaleido) manage to explicitly represent conflicts between values (Sorensen et al. 2024). Given a description of a situation (e.g. "Telling a lie to protect a friend's feelings"), Kaleido begins by exploratively generating one-hundred normative considerations (e.g. "Honesty," "Friend's well-being") before filtering them according to their relevance to the situation. It then removes repetitive items based on textual similarity, and computes relevance and valence scores for each of the remaining normative considerations (where the relevance is some number between 0 and 1, and the valence is *support*, *oppose*, or *either, depending on context*). Finally, it generates a post-hoc rationale explaining why each of the normative considerations bears on the situation (e.g. "If you value honesty, it may be better to tell the truth even if it hurts feelings").

What an AI model along these lines *can* do is to explicitly acknowledge and advert to the conflicting values implicit in its training data. This can be a valuable form of assistance, especially if it reminds one of the *variety* of values that bear on the appraisal of a situation. Even now, LLMs are much more effective than humans at exploratively overgenerating potentially relevant considerations that can then be filtered by relevance. This form of assistance is apt to draw one's attention to relevant aspects one had not yet thought of considering.

But what even a pluralist AI model such as Kaleido *cannot* do is to overcome the limitation imposed by the asystematicity of normative truth on its capacity to move beyond its training data. On the pluralist picture, even an LLM trained to acknowledge the reality of value conflicts will not be able to overcome omissions and inaccuracies in the training data *by leveraging the systematicity of truth* any more than the GPT-4 model it is based on. Insofar as the truth in normative domains is asystematic, this will rob both

kinds of models of their capacity to rely on the consistency and cohesiveness of the truth to work towards comprehensiveness. Their comprehensiveness is thus entirely reliant on the comprehensiveness of the dataset.

4. The Less Systematicity, the More Human Agency

In this final section, I shall argue that the less AI can rely on the systematicity of truth, the less we can rely on AI to do our practical deliberation for us. This is because the less the truth in normative domains is systematic, the more of a role there is for human agency and individuality in practical deliberation.

To see this, consider how the contrast between the systematicity of empirical domains on the one hand and the asystematicity of normative domains on the other produces a corresponding contrast in the structure of theoretical and practical deliberation (i.e. deliberation about what to believe and deliberation about what to do).

When deliberating about what to believe about some systematic domain such as geography, my belief-formation aims at a set of truths that are consistent and coherent, and what I end up truly believing must be consistent and cohere with what others end up truly believing. In other words, the systematicity of truth makes it the case that what *I* should believe is the same as what *anyone* should believe. The conclusion that *I* should believe that Paris is west of Prague then feels *derivative*, following as it does from a more general truth, namely that *anyone* should believe that Paris is west of Prague. In that sense, the deliberation is only *incidentally mine*.

When deliberating about what I should *do*, by contrast, the equivalence between what *I* should do and what *anyone* should do breaks down.[13] The more normative truths conflict, presenting us with considerations that pull in different directions while

[13] As pointed out notably by Williams (1985, 76–77; 1995a, 123–125; 1995b, 170), whose argument I develop and apply to the contrast between systematic and asystematic domains here.

remaining irreducibly distinct, incompatible, and incommensurable, the more *judgements of importance* will be required to determine which consideration should prevail over which in a given situation of practical deliberation.

These judgements of importance cannot be outsourced to an algorithm. One might see an attempt to do so in the "relevance scores" that Kaleido ascribes to values as a way of assessing to what extent they bear on a situation. But these relevance scores remain crucially different from the judgements of importance I have in mind. As the developers of Kaleido note, their model's training data is synthetic, meaning that it is not trained to predict whether *humans* would find a value relevant; rather, "the model's training objective is in fact closer to predicting whether a given value was likely to be generated for a particular situation by GPT-4" (Sorensen et al. 2024, 19942). This means that the relevance score captures no more than the likelihood of a certain text string figuring in text generated by GPT-4 in response to the description of a situation. The relevance score thus measures the *statistical* relevance of a type of consideration to a type of situation. But a particular consideration's importance to the agent in a particular situation goes significantly beyond such merely statistical relevance.

One might think that the problem with statistical relevance is simply that it fails to be normative, since what we really want to know is which considerations are *normatively* relevant to the situation at hand. This is true as far as it goes, but the point about importance goes even further. Normative relevance is something that a more sophisticated pluralist AI model could in principle approximate by becoming a reliable predictor of what humans would deem normatively relevant (the crucial question of *which* humans can at least partly be addressed by training the model on the judgements of the people who would go on to use the model, thereby ensuring that the model mirrors its users' judgements of normative relevance back at them).

However, even such an AI model that reliably tracked normative relevance would still fail to get at what was important *to the agent* in a particular situation. That judgement is

one that necessarily falls to the agent, and cannot be offloaded onto anyone else. This fact is obscured by an ambiguity in the question "What should I do?", which produces the impression that I might have the question answered on my behalf by someone else, or even by an AI model. But we have to distinguish two kinds of "should": the "should" that figures in the *impersonal* "What should I do?"-question and the "should" that figures in the *first-personal* "What should I do?"-question.[14] The impersonal "What should I do?"-question coincides with the "What is to be done?"-question, which asks for the recommendation of a course of action in light of all the normatively relevant considerations. But even if considering all the normatively relevant considerations yields a clear answer to the impersonal "What should I do?"-question, there still remains a *further* question for the agent who is to act on that answer: what, given all that, should *I* now do in this particular situation? This is not simply a repetition of the earlier question. And even if my answer to that further question coincides with the answer to the former question, this will not simply be a repetition of the answer. It will be an expression of the agent's judgement that what normative considerations suggest is to be done is indeed what he or she should do. This becomes evident if we consider a case in which normative considerations clearly indicate that I should ϕ, but I would very much like to ψ, and I cannot do both.[15] If I then ask myself: "What should I do?", I am not asking what course of action is favoured by the relevant normative considerations. I already know that. I am asking myself whether I really should ϕ, as the relevant normative considerations suggest I should, or whether I should follow my inclination to ψ instead.

We must accordingly be careful not to identify the "should" that figures in statements of what course of action normatively relevant considerations recommend with the

[14] Here, I draw out the consequences for AI of Williams's suggestion that practical thought is "radically first-personal" (Williams 1985, 23).

[15] This adapts an argument offered in Williams (1973, 183–185) to distinguish two kinds of *ought*; see also Williams (1995a, 123–125).

"should" in which the agent's own practical deliberation must ultimately issue. The former is only incidentally first-personal, and can equally well be answered in the third person ("What he should do is …"). The latter, however, is essentially first-personal, and can only be answered in the first person.

Suppose a sophisticated LLM trained to track what humans deem normatively relevant is used to answer a practical question. If the question really is a practical question, i.e. a question about what to *do*, there will have to be, at the end of the line, an agent *A* who enacts the answer to the question. For *A* to enact the answer, however, it is not sufficient that the AI model thinks—or its output sentences assert—that *A* should ϕ. It has to be the case that *A herself* concludes that she should ϕ, in the irreducibly first-personal sense of "should." This requires that ϕ-ing should make sense *to A* in terms of *her* judgement of what is most important to her in this situation. The practical question of what *A* should do, even if satisfactorily answered by the AI model in terms of all the normatively relevant considerations, still issues, at the end of the line, in a first-personal question for *A* that can only be answered in terms of *A*'s own judgements of importance.

Might *A* not leapfrog this first-personal question by making it an overriding principle of hers to enact whatever the AI model identifies as the thing to be done? Hardly, for even then, *A* still has to form practical conclusions in answer to first-personal practical questions. If, for instance, she has the thought "The AI said that ϕ-ing was the thing to be done," she still has to confront questions such as "Should I ϕ *now* or *later*?" and "Should I ϕ *in this way* or *in that way*?" Practical deliberation has an irreducibly first-personal dimension that expresses itself in the enduring availability of this further sense of "should." And parallel arguments can be run to distinguish two kinds of "shall" and two kinds of "ought" (as in "What shall I do?" and "What ought I to do?").[16]

[16] See Williams (1973, 183–185; 1995a, 123–125).

Of course, if this is right, then there would be a first-personal version of the "What should I do?"-question even if all normative truths could be systematically integrated into a harmonious whole.[17] But, by increasing the conflictual character of practical deliberation, the asystematicity of normative domains *accentuates and amplifies* the role of the agent. The more inevitable trade-offs, uncomfortable binds, true dilemmas, and tragic choices the normative landscape confronts us with, the more judgements are required about what is more important in a given situation.

The pluralist picture thereby not only renders the radically first-personal character of practical deliberation *salient* where the monist picture *occludes* it by suggesting that there are no irreducible conflicts left for the individual to resolve once one has recognized how everything fits together; it also gives a greater and more active role to individual agents by requiring them to determine how conflicts of values are to be navigated.

We might put this by saying that while all practical deliberation is to some degree first-personal regardless of the systematicity of truth, the asystematicity of the truth in normative domains *adds* to this first-personal dimension by allocating a *greater* role to the agent's judgements of importance. For, in navigating a conflictual normative landscape, I have to rely to a greater extent on my judgements of what strikes me as most important in a particular situation.

This comes out well in what Ruth Chang has called the "hard choices" (2017) that turn out to be ubiquitous once we recognize that normativity is *tetrachotomous* rather than trichotomous in structure—two items can be normatively related in *four* rather than three different ways: (i) one can be *better* than the other; (ii) one can be *worse* than the other; (iii) they can be *equally good*, so that one may as well flip a coin; and (iv) they can be *on*

[17] Indeed, even questions about systematically integrated non-normative domains—such as "Did Keats die in Rome?"—have first-personal analogues: "I wonder if Keats died in Rome," or "What should I believe about whether Keats died in Rome?" But these first-personal forms of the question are first-personal only incidentally. One could equally well ask "What should anyone believe about whether Keats died in Rome?"

a par, i.e. incommensurable, yet in the same neighbourhood of value. Hard choices are choices between such options that are on a par, and remain on a par even once all the relevant information is in. One choice is better in some respects, while the other is better in other respects. This does not mean that we might as well flip a coin, however. The decision can still be rational in the sense of being grounded in reasons rather than arbitrary. But coming to a rational decision requires the agent to go beyond the passive role of registering the relevance of independently given normative considerations. The agent has to play a more active role in the decision and consider which aspects are more important *to him or her*. In reminding us that even commonplace instances of practical deliberation are essentially dependent on input from the agent whose action is at issue, hard choices encourage "a fundamental shift in our understanding of what it is to be a rational agent, one that puts active, creative human agency at the center of rational thought and action" (Chang 2023, 173).

Imagine, for instance, that I am deliberating over whether I should pursue a career in philosophy or a career in consulting. Even once all the relevant normative considerations for and against each of these career choices have been exhaustively listed and carefully considered, there still remains the question of what *I* should do, especially if the various considerations do not all harmoniously assemble into an unequivocal answer. Chang presents such choices in markedly voluntaristic terms, as answerable primarily to the agent's will—on her account, it is by being willing to *commit* one way or the other that I *create* the reasons that make the choice rational.[18]

But one can also think of the process of determining what one should do as having the character of a *discovery* about oneself—and one, crucially, that only the agent him- or herself can make. Coming to a decision forces me to ask not just which considerations strike me as more important, but which considerations are more important *to me*. The

[18] See Chang (2002, 2009, 2013, 2016). Drawing on Chang's work, Goodman (2021) argues that the existence of hard choices imposes limits on how much practical reasoning AI models can do on our behalf.

decision is not merely *first*-personal in the way that every practical decision ultimately must be, but, as we naturally put it, *personal*. In coming to the conclusion that *I* should pursue a career in philosophy, because certain considerations favouring that choice are particularly important to me, I then express something distinctive of myself—something which may already have been fully formed before the process of deliberation, or which may have assumed a determinate form only through that process, but which nevertheless presents itself to me not as an expression of my will, but as a discovery about myself. Though the decision should of course still be informed by the relevant impersonal normative considerations, it should not *just* be responsive to them, but should also be true to who I am, or discover myself to have become. We might say that the form of truthfulness involved is two-faced: it encompasses being true to oneself as well as being true to the normative facts. In other words, the decision involves a demand for authenticity as well as for responsiveness to impersonal reasons.

As a result of this demand for authenticity, the conclusion that *I* should pursue a career in philosophy does not feel derivative, because it does not follow from the more general thought that *anyone* should pursue a career in philosophy. The deliberation is not just incidentally, but *essentially mine*. Practical deliberation from the point of view of a quasi-omniscient AI cannot be a substitute for this.[19] As Williams remarks: "my life, my action, is quite irreducibly mine, and to require that it is at best a *derivative* conclusion that it should be lived from the perspective that happens to be mine is an extraordinary misunderstanding" (1995b, 170). Far from being answerable only to impersonal normative considerations that an AI might weigh against each other just as well or even better than a human agent, practical deliberation systematically possesses a *first-personal* dimension, and sometimes even a *personal* dimension, in virtue of which the decision cannot be outsourced to anything or anyone else, but is essentially the agent's own.

[19] For a complementary argument why we have reasons not to want an AI that lets one know who one is and what one should do, see Leuenberger (2024).

AI models built on the assumption that practical reasoning is impersonal, and that the question "What should I do?" is equivalent to the question "What should anyone do?," neglect these first-personal and personal dimensions of practical deliberation. This remains true even if, like Kaleido, a model takes the plurality and incompatibility of values into account. For its conclusion will still take an impersonal form: it will state that, in a situation in which such-and-such conflicting values are likely to be relevant, ϕ-ing is the thing to do. If we are to be true to the first-personal and personal nature of practical thought, however, this can at most be advisory input to the agent's deliberation. The conclusion as to what the agent should actually do has to be reached *by the agent whose practical decision it is*. This vindicates the pluralist conviction that "practical decision could not in principle be made completely algorithmic, and … a conception of practical reason which aims at an algorithmic ideal must be mistaken" (Berlin and Williams 1994, 307).

The upshot is that the less systematicity normative domains exhibit, the less AI can rely on the systematicity of truth, and the less we can rely on AI to do our practical deliberation for us. There is correspondingly *more* of a role for human agency and individuality in navigating conflicts of values and making hard choices. The less systematicity, the more human agency.

5. Conclusion

Insofar as normative domains exhibit asystematicity, Amodei's optimism thus looks ill-founded: LLMs cannot necessarily rely on truths forming a systematic web across the board to self-complete and self-correct. Other ways for LLMs to move beyond their training data may yet emerge. But if the pluralist picture of normative domains is correct, LLMs cannot just rely on the systematic harmony of the landscape they strive to map, for that landscape may well turn out to be messier and more tension-ridden than expected.

And the more this is the case, the less we can rely on LLMs to do our practical deliberation for us. The more asystematic normative domains are, the more judgements of importance by the agent are called for, and these judgements of importance express and underscore first-personal and personal dimensions of practical deliberation that cannot be outsourced to AI models. Sometimes, the point is not that a decision should be absolutely and objectively the best one, but that it should be *ours*.

Bibliography

Abela, Paul. 2006. 'The Demands of Systematicity: Rational Judgment and the Structure of Nature'. In *A Companion to Kant*. Edited by Graham Bird, 408–22. Oxford: Blackwell.

Amodei, Dario. 2024. 'What if Dario Amodei Is Right About A.I.?'. Interview by Ezra Klein. *The Ezra Klein Show*, New York Times Opinion, April 12, 2024. https://www.nytimes.com/2024/04/12/opinion/ezra-klein-podcast-dario-amodei.html

Berlin, Isaiah. 2002. 'Two Concepts of Liberty'. In *Liberty*. Edited by Henry Hardy, 166–217. Oxford: Oxford University Press.

Berlin, Isaiah. 2013. 'The Pursuit of the Ideal'. In *The Crooked Timber of Humanity: Chapters in the History of Ideas*. Edited by Henry Hardy, 1–20. Princeton: Princeton University Press.

Berlin, Isaiah, and Bernard Williams. 1994. 'Pluralism and Liberalism: A Reply'. *Political Studies* 42 (2): 306–309.

Blum, Christian. 2023. 'Value Pluralism versus Value Monism'. *Acta Analytica* 38 (4): 627–652.

Brooks, Thom, and Sebastian Stein, eds. 2017. *Hegel's Political Philosophy: On the Normative Significance of Method and System*. Oxford: Oxford University Press.

Chang, Ruth, ed. 1997a. *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA: Harvard University Press.

Chang, Ruth. 1997b. 'Introduction'. In *Incommensurability, Incomparability, and Practical Reason*. Edited by Ruth Chang, 1–34. Cambridge, MA: Harvard University Press.

Chang, Ruth. 2002. *Making Comparisons Count*. London: Routledge.

Chang, Ruth. 2009. 'Voluntarist Reasons and the Sources of Normativity'. In *Reasons for Action*. Edited by David Sobel and Steven Wall, 243–271. Cambridge: Cambridge University Press.

Chang, Ruth. 2013. 'Commitments, Reasons, and the Will'. In *Oxford Studies in Metaethics, Volume 8*. Edited by Russ Shafer-Landau, 74–113. Oxford: Oxford University Press.

Chang, Ruth. 2015a. 'Value Incomparability and Incommensurability'. In *The Oxford Handbook of Value Theory*. Edited by Iwao Hirose and Jonas Olson, 205–224. Oxford: Oxford University Press.

Chang, Ruth. 2015b. 'Value Pluralism'. In *International Encyclopedia of the Social & Behavioral Sciences*. Edited by James D. Wright. Vol. 25, 21–26. Oxford: Elsevier.

Chang, Ruth. 2016. 'Comparativism: The Grounds of Rational Choice'. In *Weighing Reasons*. Edited by Errol Lord and Barry Maguire, 213–240. New York: Oxford University Press.

Chang, Ruth. 2017. 'Hard Choices'. *Journal of the American Philosophical Association* 3 (1): 1–21.

Chang, Ruth. 2023. 'Three Dogmas of Normativity'. *Journal of Applied Philosophy* 40 (2): 173–204.

Chappell, Timothy. 2009. *Ethics and Experience: Life Beyond Moral Theory*. Durham: Acumen.

Cueni, Damian. 2024. 'Constructing Liberty and Equality – Political, Not Juridical'. *Jurisprudence* 15 (3): 341–360.

Cummins, Robert, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz. 2001. 'Systematicity and the Cognition of Structured Domains'. *Journal of Philosophy* 98 (4): 167–185.

Dancy, Jonathan. 2004. *Ethics without Principles*. Oxford: Clarendon Press.

Feng, Nick, Lina Marsso, S. Getir Yaman, Isobel Standen, Yesugen Baatartogtokh, Reem Ayad, Victória Oldemburgo de Mello, Bev Townsend, Hanne Bartels, Ana Cavalcanti, Radu Calinescu, and Marsha Chechik. 2024. 'Normative Requirements Operationalization with Large Language Models'. *arXiv*.

Franks, Paul W. 2005. *All or Nothing: Systematicity, Transcendental Arguments, and Skepticism in German Idealism*. Cambridge, MA: Harvard University Press.

Goodman, Bryce. 2021. "Hard Choices and Hard Limits for Artificial Intelligence." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.

Greene, Joshua. 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Press.

Greene, Joshua D. 2016. 'Our Driverless Dilemma'. *Science* 352 (6293): 1514–1515.

Guyer, Paul. 2003. 'Kant on the Systematicity of Nature: Two Puzzles'. *History of Philosophy Quarterly* 20 (3): 277–295.

Guyer, Paul. 2005. *Kant's System of Nature and Freedom*. Oxford: Oxford University Press.

Hämäläinen, Nora. 2009. 'Is Moral Theory Harmful in Practice?—Relocating Anti-theory in Contemporary Ethics'. *Ethical Theory and Moral Practice* 12 (5): 539–553.

Heathwood, Chris. 2015. 'Monism and Pluralism about Value'. In *The Oxford Handbook of Value Theory*. Edited by Iwao Hirose and Jonas Olson, 136–157. Oxford: Oxford University Press.

Kambartel, Friedrich. 1969. "System' und 'Begründung' als wissenschaftliche und philosophische Ordnungsbegriffe bei und vor Kant'. In *Philosophie und Rechtswissenschaft: Zum Problem ihrer Beziehung im 19. Jahrhundert*. Edited by Jürgen Blühdorn and Joachim Ritter, 99–122. Frankfurt am Main: Klostermann.

Kekes, John. 1993. *The Morality of Pluralism*. Princeton: Princeton University Press.

Kitcher, Philip. 1986. 'Projecting the Order of Nature'. In *Kant's Philosophy of Material Nature*. Edited by Robert Butts, 201–235. Boston: D. Reidel.

Kretzmann, Norman, and Eleonore Stump. 1989. *The Cambridge Translations of Medieval Philosophical Texts: Volume 1, Logic and the Philosophy of Language*. Cambridge: Cambridge University Press.

Larmore, Charles. 1987. *Patterns of Moral Complexity*. Cambridge: Cambridge University Press.

Leuenberger, Muriel. 2024. 'Should You Let AI Tell You Who You Are and What You Should Do?'. In *AI Morality*. Edited by David Edmonds, 160–69. Oxford: Oxford University Press.

Losano, Mario G. 1968. *Sistema e struttura nel diritto, vol. 1: Dalle origini alla scuola storica*. Turin: Giuffrè.

MacIntyre, Alasdair C. 2007. *After Virtue: A Study in Moral Theory*. 3rd ed. Notre Dame: University of Notre Dame Press.

Mason, Elinor. 2023. 'Value Pluralism'. In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Summer 2023 ed.

Messer, August. 1907. 'Besprechung von Otto Ritschl: System und systematische Methode in der Geschichte des wissenschaftlichen Sprachgebrauchs und der philosophischen Methodologie'. *Göttinger gelehrte Anzeigen* 169 (8): 659–66.

Millgram, Elijah, and Paul Thagard. 1996. 'Deliberative Coherence'. *Synthese* 108 (1): 63–88.

Nagel, Thomas. 2001. 'Pluralism and Coherence'. In *The Legacy of Isaiah Berlin*. Edited by Mark Lilla, Ronald Dworkin and Robert Silvers, 105–111. New York: New York Review of Books.

Queloz, Matthieu. 2024. 'The Dworkin–Williams Debate: Liberty, Conceptual Integrity, and Tragic Conflict in Politics'. *Philosophy and Phenomenological Research* 109 (1): 3–29.

Rescher, Nicholas. 1979. *Cognitive Systematization: A Systems Theoretic Approach to a Coherentist Theory of Knowledge*. Oxford: Blackwell.

Rescher, Nicholas. 1981. 'Leibniz and the Concept of a System'. In *Leibniz's Metaphysics of Nature: A Group of Essays*, 29–41. Dordrecht: Springer.

Rescher, Nicholas. 2000. *Kant and the Reach of Reason: Studies in Kant's Theory of Rational Systematization*. Cambridge: Cambridge University Press.

Rescher, Nicholas. 2005. *Cognitive Harmony: The Role of Systemic Harmony in the Constitution of Knowledge*. Pittsburgh, PA: University of Pittsburgh Press.

Ritschl, Otto. 1906. *System und systematische Methode in der Geschichte des wissenschaftlichen Sprachgebrauchs und der philosophischen Methodologie*. Bonn: C. Georgi.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Viking.

Sen, Amartya. 1981. 'Plural Utility'. *Proceedings of the Aristotelian Society* 81 (1): 193–216.

Sorensen, Taylor, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024. "Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties." Proceedings of the AAAI Conference on Artificial Intelligence.

Stein, Aloys von der. 1968. 'Der Systembegriff in seiner geschichtlichen Entwicklung'. In *System und Klassifikation in Wissenschaft und Dokumentation*. Edited by Alwin Diemer, 1–13. Meisenheim am Glan: A. Hain.

Stocker, Michael. 1990. *Plural and Conflicting Values*. Oxford: Clarendon Press.

Tasioulas, John. 2022. 'Artificial Intelligence, Humanistic Ethics'. *Daedalus* 151 (2): 232–243.

Thompson, Kevin. 2017. 'Systematicity and Normative Justification: The Method of Hegel's Philosophical Science of Right '. In *Hegel's Political Philosophy: On the Normative Significance of Method and System*. Edited by Thom Brooks and Sebastian Stein, 44–66. Oxford: Oxford University Press.

Troje, Hans Erich. 1969. 'Wissenschaftlichkeit und System in der Jurisprudenz des 16. Jahrhunderts'. In *Philosophie und Rechtswissenschaft: Zum Problem ihrer Beziehung im 19.*

Jahrhundert. Edited by Jürgen Blühdorn and Joachim Ritter, 63–88. Frankfurt am Main: Klostermann.

Vieillard-Baron, Jean-Louis. 1975. 'Le concept de système de Leibniz à Condillac'. In *Akten des II. Internationalen Leibniz-Kongresses Hannover, 17.-22. Juli 1972*. Edited by Kurt Müller, Heinrich Schepers and Wilhelm Totok, 97–103. Wiesbaden: F. Steiner.

Williams, Bernard. 1973. 'Ethical Consistency'. In *Problems of the Self*, 166-186. Cambridge: Cambridge University Press.

Williams, Bernard. 1981a. 'Conflicts of Values'. In *Moral Luck*, 71–82. Cambridge: Cambridge University Press.

Williams, Bernard. 1981b. 'Moral Luck'. In *Moral Luck*, 20–39. Cambridge: Cambridge University Press.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Routledge Classics Edition. London: Routledge.

Williams, Bernard. 1995a. 'Formal and Substantial Individualism'. In *Making Sense of Humanity and Other Philosophical Papers 1982–1993*, 123–34. Cambridge: Cambridge University Press.

Williams, Bernard. 1995b. 'The Point of View of the Universe: Sidgwick and the Ambitions of Ethics'. In *Making Sense of Humanity and Other Philosophical Papers 1982–1993*, 153–71. Cambridge: Cambridge University Press.

Williams, Bernard. 1995c. 'What Does Intuitionism Imply?'. In *Making Sense of Humanity and Other Philosophical Papers 1982–1993*, 182–191. Cambridge: Cambridge University Press.

Williams, Bernard. 2001. *Morality: An Introduction to Ethics*. Cambridge: Cambridge University Press.

Williams, Bernard. 2005a. 'From Freedom to Liberty: The Construction of a Political Value'. In *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Edited by Geoffrey Hawthorne, 75–96. Princeton: Princeton University Press.

Williams, Bernard. 2005b. 'Pluralism, Community and Left Wittgensteinianism'. In *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Edited by Geoffrey Hawthorne, 29–39. Princeton: Princeton University Press.

Williams, Bernard. 2013. 'Introduction'. In *Concepts and Categories: Philosophical Essays*. Edited by Henry Hardy. 2nd ed, xxix–xxxix. Princeton: Princeton University Press.

Ypi, Lea. 2021. *The Architectonic of Reason: Purposiveness and Systematic Unity in Kant's Critique of Pure Reason*. Oxford: Oxford University Press.