

Iterative Explication

Catherine Z. Elgin

Abstract

Carnap maintains that explication refines and sharpens everyday concepts, enabling them to perform rigorous philosophical and scientific functions. Building on Carnap and Chang, I argue that conceptual enhancement often involves iterative explication. Each step is an improvement on its predecessor and sets the stage for further conceptual refinement. I draw on familiar examples from the history of philosophy to illustrate the procedure and its benefits.

The Problem

Familiar, widely endorsed epistemological theories apparently do not accommodate philosophy itself. It is doubtful that our favored philosophical positions are reliable, justified, or virtuously generated and/or sustained. It is doubtful that any of them is true as it stands. Still, we seem to be doing something epistemically reputable. If so, epistemology ought to be able to show how philosophy itself affords understanding and, ideally, how it advances understanding. Otherwise, our pretensions to prescribe standards for epistemic acceptability seem hypocritical and self-undermining. Quine maintains that philosophy is continuous with natural science (1969:83). I'd argue that it is continuous with inquiry of all sorts.

The Aim

'The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term.' So says Sellars (1963:1). On the assumption that something close to this captures what we are trying to do, I will argue (1) that we *do* make advances in philosophy and (2) that they are in no small measure products of epistemic iteration – a procedure by which we fine tune our concepts, methods,

commitments, and practices by a process of mutual accommodation where we test them against one another and revise them accordingly. In this, I will suggest philosophy is like other disciplines.

To set the stage, I need to say something about how capacious *the broadest possible sense of the term* is. The ‘things’ philosophy investigates include the material and the immaterial – objects, properties, relations; numbers, sets, norms, methods, and standards; necessities, possibilities, and impossibilities; rules and violations; actions and omissions; existents and non-existents; the extensional, intensional, and hyperintensional; spaces, times, and space-times; minds and bodies; enduring objects and ephemeral events. The list goes on. Nothing imaginable or unimaginable, intelligible or unintelligible is left out. It is a lot to aspire to understand. That being so, philosophy might, not unreasonably, be charged with overweening ambition. Perhaps we are guilty. But given our propensity to take the remotest counterexample seriously, to concern ourselves with what is the case in distant possible (or even impossible) worlds, to worry about malevolent demons and brains in vats, to concern ourselves with the fates of imaginary cave dwellers and imaginary zombies, the charge seems apt. We *do* think that our account should yield correct verdicts on wildly improbable, clearly contrived cases. We are convinced that it is imperative to take such cases seriously.

We seek to understand not just the things, but how they hang together – how even seemingly distant philosophical truths are connected. No matter how firmly established they were, we would not be satisfied with a list of individual truths:

To be is to be the value of a variable.

Monads have no windows.

Justice is fairness.

We should call no one happy until he is dead.

Ought implies *can*.

Even if we were convinced of the truth of each entry on the list, we would still want to understand how they relate to each other. As Sellars indicates, we are willing to at least start out pretty open-minded about what hanging together amounts to; but some sort of robust relation – perhaps dependence or interdependence, perhaps supervenience or grounding, perhaps inferential, evidential, or probabilistic connection – is called for. Merely conjoining the truths on the list would not be good enough.

Luckily, there is a division of cognitive labor. None of us is expected to weave the whole multi-dimensional philosophical tapestry on her own. Each of us works on her own small patch. Nevertheless, we can't completely ignore the rest of the tapestry. We have to consider whether and how the commitments about our own small patch fit into the big picture and advance the overall project. In devoting her professional life to delineating the relation between freedom and responsibility, Sarah has to keep a weather eye out to ensure that her conception of freedom does not conflict with the laws of nature. In devoting his professional life to arguing for aesthetic cognitivism, Harry has to be on guard that his account of aesthetic truth does not commit him to the liar paradox. Because the things philosophy investigates are supposed to hang together, findings from distant realms may have consequences that bear on the tenability of local theories.

The vastness of the project is daunting. It raises questions about the adequacy of our conceptual, inferential, and investigatory tools. Wittgenstein thinks there's no worry. 'All the propositions of our everyday language, just as they stand, are in perfect logical order' (1961:§5.5563). If so, conceptual analysis of everyday language will show how available concepts suffice for philosophical purposes. But everyday terms, and the concepts they express, are often vague or imprecise. An accurate analysis would just replicate the vagueness or imprecision. Since,

for example, the boundaries of the concept of happiness are vague, an accurate conceptual analysis of *happiness* should reflect that vagueness. If we want precision – if, for example, we want to determine whether a mildly favorable mood or a momentary flash of euphoria qualifies as happiness – we need something else. Moreover, since terms in ordinary language do not answer to sharp criteria, translation from one language to another is often inexact.

Wittgenstein's optimism is, I think, misguided. Whatever we think about the adequacy of ordinary language for ordinary purposes, conceptual analysis will not deliver what philosophy needs. Philosophy seeks precision; it demands rigor. Since everyday concepts are imprecise, everyday reasoning is loose, and everyday means are not always suited to philosophical ends, everyday language and logic, whatever their merits, are not up to the task. Regimentation is required.

This should not be a surprise. Rarely, if ever, do we formulate concepts by specifying necessary and sufficient conditions. We coin terms and formulate concepts to do a particular job – to enable us to draw the distinctions that we want to draw. We don't worry much about borderline cases if we expect to keep our distance from the borders. In everyday contexts, where we want to distinguish cars from trucks, or apples from pears, such a practice is effective. There aren't all that many borderline cases, and they don't make much difference. But philosophy, science, and law often require that we increase the precision, exactitude, and rigor of our conceptual and inferential resources. So rather than thinking that our everyday concepts and modes of reasoning are up to the task that philosophy wants to use them for, I suggest, we should think (and hope) that they are susceptible of improvement so that their successors can do a better job. This is not always just a matter of eliminating vagueness or imprecision. Sometimes a familiar

concept, even if it has sharp enough edges, is not quite suited to the philosophical purposes we want to use it for.

Explication

It might seem that the solution to our difficulty is to replace a vague or imprecise, or unsuitable term with one that has necessary and sufficient conditions. Although we are pretty good at distinguishing living things from non-living things, we may, for example, be in a quandary about viruses. They self-replicate like quintessentially living things; but they do so only within and by exploiting the cellular resources of living hosts, whereas quintessentially living things rely on their own cellular resources. Are they alive or not? We want a formula:

$$(x) x \text{ is alive} \equiv y$$

that would settle the question.

Although terms in ordinary language lack necessary and sufficient conditions, we could surely introduce stipulative definitions for our technical terms. So we could provide a definition of ‘alive’ from which it immediately follows that a virus is alive or that it is not. But we do not want an arbitrary definition. When we ask whether a virus is alive, the concept of *alive* that figures in our question should be if not the same then at least close to the one we use to characterize quintessentially living things. Our answer should be suitably connected to our question. Are viruses enough like things that we already consider alive in the respects that qualify as alive? We can’t stipulate an answer to this question. This is why, Carnap maintained, we should resort to explication. Unlike a stipulative definition, an explicandum is tethered to antecedently accepted uses of a term. Any explication of *alive* should agree with antecedent verdicts about many, if not all, non-disputed instances. A measure of flexibility is allowed. If, for example, an explication

aligned with relatively few, but manifestly central instances, that might be acceptable. But tethering the explanandum to antecedently accepted usage is mandatory.

Still, there's a problem. Even if we could formulate a definition that was suitably tethered, there is no assurance that it would articulate conditions whose satisfaction we can ascertain. If, because of limits on our measuring devices or modes of detection or breadth of imagination, we cannot tell whether y obtains, the news that

$$(x) x \text{ is alive} \equiv y$$

is not likely to foster inquiry or promote the achievement of our epistemic ends. If, for example, we agreed with Aristotle's claim in *De Anima* that

$$(x) x \text{ is alive} \equiv x \text{ has a soul},$$

that wouldn't help. For we have no way to tell – indeed no idea how to tell – whether a virus (or, for that matter, anything else) has a soul.

Moreover, our explication should promote our interests. If our aim is to advance biological science, whether or not a virus has a soul may be irrelevant. There is, as Carnap recognized, a pragmatic dimension to explication. An adequate explication must be suited to its epistemic purpose.

Because an explication can and typically does diverge from standard usage, a single pre-theoretic term can be explicated in multiple ways. A term like 'alive' admits of divergent explications, each fostering a different line of inquiry. One might, for example, figure in inquiries into whether we should classify viruses as alive, hence be relevant to biology. Another might figure in inquiries into whether brain-dead organ transplant donors are alive when their organs are mechanically supplied with blood to keep them viable; that would bear on biomedical ethics. Since

the goal of explication is to craft concepts that will facilitate inquiry, the choice of modifications to available concepts is guided by the goals of the inquiry (see Carnap 1950, Brun 2016).

Let's switch examples: Suppose the goal of our inquiry is to figure out whether eudaimonia is the highest human good. Immediately we have a problem. There seems to be no English equivalent of the Greek word. It is sometimes translated as 'happiness', but eudaimonia is not a feeling or a state of mind. A better translation is 'flourishing'. But it isn't quite right. The ordinary language concept of flourishing means something like well-being. If our inquiry concerns organismic biology, a generic concept of well-being might be what we seek. It would apply to healthy animals and plants in propitious environments. That is one route an explication of flourishing could reasonably take. But flourishing as generic well-being is too broad for our purposes. Healthy plants and animals flourish when ecological conditions are apt. So do we. Nevertheless, the feature we share with these other organisms seems not to be a candidate for the highest (distinctively) *human* good. Nor can we simply throw in a constraint, so that the flourishing that we are interested in is simply the well-being of healthy *human* organisms in biologically propitious environments. That would restrict our focus to human biology. But we are not mainly interested in biological well-being – the condition of hale, hardy, energetic human animals. We want a concept that not only integrates, but also emphasizes the importance of character to human well-being.

We need to craft a concept out of available resources, by pruning, padding, and reconfiguring as required. We sharpen the concept's edges – eliminating, as far as possible, vagueness, imprecision, and indeterminacy. We exclude or sideline unwanted commitments. We focus on the end in view. We reconfigure the concept of flourishing, crafting a quasi-technical term to serve our philosophical purpose. We augment it and adjust our focus, insisting that human well-being

involves more than the animal well-being of *Homo sapiens*. The resulting concept is still tethered in the everyday concept, but it is considerably sharper, more focused, and more restricted in its scope. These divergences serve our epistemic purposes.

Explication is a delicate business. A classicist might seek an explication of eudaimonia that would afford an understanding of Aristotle. The goal then would be to frame a concept that illuminates the *Nicomachean Ethics* and perhaps sheds light of some of Aristotle's other works. A neo-Aristotelian might seek an explication that would render the position plausible for modern audiences. It might diverge (at least a bit) from one that simply sought to make sense of the *Ethics*. It might, for example, eliminate the conviction that magnanimity is required for flourishing, since that seems to imply that only the wealthy can flourish. The neo-Aristotelian might go on to replace magnanimity with generosity, since generosity is a virtue that even those of limited means can exercise. The goal of such explications would be to afford concepts that enable us to express theories that their adherents are already committed to. In such cases, their adherents have pretty much figured out where they stand, but need to hone their terminology to express the commitment to eudaimonia that they have already made.

I want to focus on an earlier stage, where the goal is to fix the boundaries of a concept so as to *facilitate* inquiry. It is possible and often desirable to mark out the contours of a concept independently of and prior to taking any stand on whether that concept has instances or whether the concept, with those instances, will perform the function we seek to perform. Rather than trying to articulate a concept that makes sense of Aristotle, or one that makes it plausible that flourishing is in fact the highest human good, at this stage we seek to craft a concept that enables us to *investigate* whether flourishing is the highest human good. We do not want to beg the question in favor of or against Aristotelianism. We do not want to craft the concept of flourishing so that it

follows analytically that flourishing is, or that it is not, the highest human good. (See S. Z. Elgin 2020). Rather, we want a concept that enables us to more precisely demarcate the problem we hope to solve. For that we need to be able to say, ‘*This* is what we are talking about. Now is *it* the highest human good?’ The concept we use will not be the everyday concept. It will be carefully crafted to enable us to investigate the matter without begging the question.

Although I started with a problem of translation, as the shift from eudaimonia to flourishing shows, the issue of explication is equally fraught when we are talking about the explication of a term in its own language. We may sharpen, smooth, extend, and distend when we attempt to adjust the parameters of a familiar concept to enable it to serve an end for which it is currently not quite suitable. In ordinary language, the word ‘flourishing’ is a term for well-being. If we seek to focus on social well-being, we might circumscribe the term so that it applies to the well-being not just of an organism as such – a dandelion or a cockroach – but to its well-being as a social animal. That still seems too broad, since it applies to gorillas, bonobos, and elephants as well as to humans. So we circumscribe further, maybe limiting our focus to the well-being of an organism as a *moral* or *political* animal – one who subjects its behavior to laws. Eventually, perhaps, we fine-tune the concept to the point that it can suit the role we want it to play in our inquiry regarding the highest human good. My point here is not to advocate for any particular explication of a concept, but to show how the process of explication can lead to conceptual refinement.

To sum up: An explication requires a *tether* that connects it to, and grounds its acceptability in, a pre-theoretical concept, which may be drawn from everyday usage, from a different theory, or from an earlier stage of theorizing about the current issue. An explication must have a *goal* – a problem in the solution of which the refined concept is designed to play a role. The *purpose* of the process of explication is to mould the tether so that it can figure in the achievement of the goal.

The fruit of an explication should be better suited to solving the problem than its pre-theoretical predecessor was.

Carnap seems to treat this as a one step process, where we replace an unsatisfactory explanans with a wholly satisfactory explanandum. But there is no requirement that this be done in a single step. We might start with an everyday concept, explicate it to yield a concept that is better suited to achieving the goal, but not yet good enough. Then we might explicate that concept, yielding further improvements. This is an iterative process, with each successive step bringing us closer to our goal.

Although Carnap focuses on the explication of concepts, the same strategy is used to fine-tune standards of evidence, norms of acceptability, and modes of inference. Through explication logicians develop modal logics to regiment our rather rough and ready everyday reasoning about possibility and necessity. Moreover, only when we have the resources of such a logic can we recognize the need for, and devise an appropriate semantics. The availability of multiple modal logics with different strengths underscores the pragmatic dimension, since suitability for a purpose is critical to deciding which logic to use.¹

If explication required only a single step – replacing an inapt term with an appropriately regimented one – the question would simply be whether a proposed reconfiguration of the concept or mode of reasoning achieves its goal. But if iteration is required, an element is missing. The iterative process needs a criterion for ‘better than’. An iterative chain of explications requires some way to tell whether a putative revision brings us closer to our goal. There is no reason to think that the same criterion will be suitable for every explication. But there are a few general features that viable candidates share. A suitable criterion should be *consistent*; it should display test/retest

¹ I am grateful to Georg Brun for this point.

agreement. If an agent were to reapply the same criterion to the same candidate in the same context, that criterion should deliver the same verdict; and if she were to apply a similar criterion to a similar candidate in similar circumstances, that criterion should deliver a similar verdict. A suitable criterion should be *applicable*; it should be such that agents can use it to arrive at a verdict. It should not require agents to make to infinitely fine or infinitely many measurements, appeal to undetectable magnitudes, or rely on tests that cannot actually be run. The criterion's verdicts should be *epistemically accessible*, so that epistemic agents, in the contexts where the criterion is to be applied, can tell whether it has been satisfied. It should be *intersubjectively deployable*, that is, it should be such that multiple epistemic agents can use it. Verdicts then cannot be wholly subjective. And its verdicts *susceptible of intersubjective agreement*. Multiple epistemic agents should be able to use the same criterion, get the same (or close enough to the same) results and agree about what those results are. Moreover, the verdicts the criterion sanctions should be *plausible* – or at least, not implausible – given the relevant antecedent commitments of the users. Users should have reason to think that the verdicts indicate that a proposed explication is, or that it is not, an improvement.

Initial criteria may be vague, imprecise, gappy. They may be capable only of supplying relatively crude verdicts. This is not a problem so long as the criteria themselves are susceptible of recognizable improvement. In that case, we can refine our criteria as we advance our understanding of the concept, the problem, and the aim.

Epistemic Iteration

In *Inventing Temperature* (2004) Hasok Chang describes the epistemically iterative process by which the thermometer (a device) and temperature (the magnitude it measures) evolved in

tandem – each facilitating refinement of the other. My account of epistemic iteration draws heavily on his, so it is worth sketching part of the process he describes.

Long before the invention of the thermometer, people recognized that they were sensitive to differences in warmth. They knew that winters are colder than summers, that water freezes when the weather is cold, melts when it gets warmer, boils away when put over a flame. Because they recognized that many of these phenomena exhibit regular patterns that are not wholly subjective, they took it that phenomenal experiences of warmth track something. And because at least some of the patterns seem important, they thought it would be epistemically worthwhile to learn what they track. They also recognized that mimicking phenomenal experiences would not yield the result they sought. The relevant subjective assessments are sometimes at odds. The wind feels cold to one person; mild to another. The lake feels chilly when one first dives in, mild a few minutes later. Still, some of the regularities people discerned – such as those that concern water’s changes of state, and those that correlate with the sequence of the seasons – seem independent of anyone’s discerning them. If the subjective factor could be factored out, the remainder would presumably be whatever mind-independent property underwrites the regularities.

Since subjective assessments of warmth are inconstant and mutually inconsistent, they are not reliable indicators of a mind-independent property. But at the outset, there seemed nothing better to go on. So rather than jettisoning them (and the project) completely, investigators exploited regularities and irregularities in the inconsistent data. They noticed that although some reports are at odds, a considerable number are in accord. There also seem to be identifiable circumstances in which many informants give mutually consistent verdicts. And there seem to be identifiable individuals whose judgments regularly and sharply diverge from those of their peers. There are then discernible patterns of agreement and disagreement. These supply evidence.

Initially, investigators set aside verdicts on which there was no consensus and disregarded reports of informants who are frequent outliers. They then focused on the remainder. Let us call these the *clear cases*. They supply the tether. A item qualifies as a clear case if and only if there is widespread agreement that it is hot or that it is cold; otherwise it is unclear. Many clear cases can be put in a linear order. It is widely agreed, let us say, p is warmer than q , and that q is warmer than r ; and that p is warmer than r . It is widely agreed that p is not warmer than itself and that when p is warmer than q , q is not warmer than p . This supplied the resources to create an ordinal sequence – a scale that enabled investigators to determine that one thing is warmer than, is as warm as, or is less warm than another. Obviously, they could not locate every assessment of warmth or even every clear case on the scale; but they did not need to. The first step was just to establish a schema that could draw and order crude but consistent distinctions. That sequence affords the conceptual grounding for a thermoscope – a device whose behavior correlates with, hence can be seen as reflective of, where things stand on this ordinal scale. Investigators created the device, relying on the knowledge that familiar liquids visibly expand when heated and contract when cooled. They validated early thermoscopes by testing whether their verdicts aligned about the clear cases. Exact alignment was not expected. Both the initial clear cases and the initial measuring devices were rough. If a thermoscope agreed pretty well with most of the clear cases, that was good enough. Because a thermoscope's column of liquid rose in cases that were considered clearly hot, and fell in cases that were considered clearly cold, it could provide a benchmark. Investigators could use the device to adjudicate unclear cases. If when the device tested an unclear case the liquid rose, that case was reclassified as clearly hot; if it fell, the item was reclassified as clearly cold. The initial thermoscopes were not expected to settle every case, but they enabled investigators reclassify some. Incorporating thermoscopic findings enabled

investigators to expand the classes of clear cases and narrow the class of unclear ones. That in turn enabled them to improve the device, so that it could make finer distinctions. Investigators could then use the improved device to adjudicate remaining unclear cases. Again it settled some, but not all. They then further augmented the classes of clear cases and used the augmented classes as a basis for yet further improvements in the thermoscope. (One improvement might, for example, have been to narrow the tube so that smaller differences in the height of the liquid would be visible.) The process could be continued indefinitely.

Chang's description aligns with the template I described earlier.

- The original *tether* consisted of pre-theoretic agreement about clear cases of hot and cold.
- The *goal* was to construct an objective, measurable magnitude – temperature.
- The *purpose* of the iterative process was to refine the tether so as to achieve the aim.
- The initial *criterion* was agreement with the readings of thermoscopes that had been calibrated against clear cases.
- *Revised tethers* resulted from augmented classes of clear cases. This set the stage for fine-tuning the thermoscopes, enabling them to adjudicate some of the remaining unclear cases.

As you would expect, the actual history is more complex than my schematic description suggests. Success was not guaranteed. Scientists had to invent the thermoscope and insure that at each stage it continued to track the evolving sequence of classes of clear cases. This was an engineering project that was constrained by the available conceptual, imaginative, and material resources. It could have failed at the outset, or have reached a limit beyond which no further iterations were possible. And failure, if it occurred could have been due either to engineering limitations or to the irresolvable murkiness of the cases that remained unsettled. Moreover, there was, at the outset, no guarantee that phenomenal assessments of warmth admitted of the sort of factor analysis that

enabled the subjective element and the objective element to be prized apart. Such an iterative process might fail wholesale; or a particular iterative step might resist further refinement. Epistemic iteration is fallible.

Conceptual Refinement

Our question is whether epistemic iteration figures in conceptual refinement in philosophy. I think it does. And if it does, we have a defense against those who contend that philosophy makes no progress. I would argue that although the questions are hard, we *have* made progress, in that we have over time refined the concepts, methods, and problems we address. John Dewey contends that the mark of a good answer is that it enables us to ask good questions (1916:104-106). If this is right, then epistemic progress consists not merely in finding answers, but in finding new and better questions to ask. And a good answer would be, not one that provides a fixed and conclusive solution to a particular question, but one that serves as a springboard for further epistemic advancement.

Conceptual refinement is called for when available concepts are inadequate to the demands placed on them. Most available concepts are not, we may suppose, entirely bankrupt. They have proven themselves to be reasonably adequate for earlier tasks. But they are vague, gappy, or not quite suitable for the use we want to make of them. Work needs to be done.

A concept has a putative domain – the class of cases for which the question of its applicability arises. In explicating a concept, we begin by identifying clear cases. Some are clear instances; others, clear foils. Clear instances are those to which the concept obviously applies; clear foils are those to which it obviously fails to apply. These provide the tether. There remain cases which fall into neither class. Our current criteria for instantiation do not enable us to say whether the concept applies or not. Our inquiry has a goal – an issue we want to address, a problem we want to solve,

or a question we want to answer. That indicates the direction we should take in making revisions. The purpose of a proposed explication is to augment the clear cases and resolve the status of the unclear ones in a way that fosters the realization of the goal. Revisions should neither random nor arbitrary. So we need a criterion, whose satisfaction would indicate that we are on the right track. That may seem to be a problem.

In the case Chang describes, investigators had to invent, build, and successively refine a device. Philosophers don't have that option. Gadgetry will not take us where we want to go. In the quest to understand causality, we might, of course, venture far enough into the engineering realm to construct devices that test for robust correlation. That's not enough. A mantra of statistics is 'Correlation does not imply causation'. But the most we can expect our devices to supply is robust correlation. So our question is what can we use for a criterion? This is where what Georg Brun (2016) calls conceptual re-engineering comes into play.

When W. D. Ross (2002) asks 'What Makes Right Acts Right?' he is looking for a criterion for moral rightness. The form of his question shows what we want in a criterion. We have clear cases to which a concept – in Ross's discussion, the concept of *right acts* – applies, and clear foils – acts which are manifestly wrong – to which it fails to apply. We want to identify the *basis* of the concept that underwrites correct classification. We want to know: what makes *x*-es *x*-es? When we can't answer that immediately, we may be in a position to answer what makes clear *x*-es clearly *x*-es and what makes clear $\sim x$ -es clearly $\sim x$ -es? We venture a hypothesis that pretty much captures the clear cases and repudiates the clear foils, and we see if, and how, it sorts the unclear cases. We recognize that our pre-theoretic partition of the domain is not exhaustive. A candidate criterion may suggest that the scope of the concept is wider or narrower than we thought and/or that it contains more complicated cases than we first imagined. And it may be that a few of the cases

turn out to be misclassified. As in the iterations that figured in the invention of temperature, we expand the class of clear cases and clear foils. We also expand the class of disputed cases.² We find ourselves entertaining possible, but not actual cases. We notice that if ϕ is the basis for something's being ψ , then hypothetical case z , being ϕ , would be ψ . Is that plausible? If so, we include z (which does not actually exist) among the clear cases. If not, we either reject the hypothesis or revise it so that it no longer includes z . Either way, the result enhances our understanding of the scope and limits of ψ .

Locke on Personal Identity

This is all horribly schematic, so let's look at Locke's (1689/1894) discussion of personal identity to see how it plays out. I want to focus not on the adequacy of Locke's position, but on the trajectory of his argument. Locke, as is well known, is interested in the enduring identity of a person – that is, what makes someone the same person over an extended period of time. He ventures several hypotheses. The first is that the identity of a person is grounded in sameness of substance. He begins with the hypothesis that is that the substance is physical. Same person = same physical object. That won't work. People grow; they gain and lose weight; they go bald, etc. So, he suggests, maybe personal identity *is* identity of substance, but the substance isn't physical. Maybe the criterion is sameness of soul. But, he says, if that is so, we still run into problems. Some believe in the transmigration of souls. We can't rule that out a priori. If souls transmigrate, then it is possible that Socrates, St. Augustine, and Caesar Borgia had the same soul. But it is not remotely plausible that we would consider them the same person. Nor if Heliogabalus was reborn as a hog, would we consider that hog a person.

² This may also have happened in the quest Chang describes. He does not say.

Notice what is going on here: Locke proposes a criterion for the identity of a person over time. It answers to the clear cases – we readily recognize our neighbor as the same person from week to week; we recognize that he is a different person from his brother. We test the criterion of sameness of physical substance and it fails. Bill is the same person even though he has lost weight. His daughter Sally is the same person even though she has grown a lot taller. We move on to sameness of soul (a related, but different criterion, since the substance in question is immaterial). This is a bit harder, since we don't have an easy way to track souls. So we can't rule it out by inspection of cases as we did sameness of physical substance. But, Locke says, it has an untenable consequence. If souls transmigrate (as some say they do) then individuals whom we regard as clearly distinct qualify as the same person. If souls transmigrate to other animals (as some say they do), then the individual who was once the person Heliogabalus is now a hog. But, we are convinced that Socrates is a different person from Cesar Borgia, and the person who is Heliogabalus is not and never will be a hog. Locke does not commit himself to the transmigration of souls. He rejects the criterion on the basis of a mere possibility. In so doing, he expands the class of clear cases to include verdicts about hypotheticals and counterfactuals. A satisfactory criterion should not just get the clear actual cases right, it should also get what we consider clear counterfactual and hypothetical cases right.

These first two attempts might seem to be a matter of incremental improvement. But iterative explication is not always a matter of incrementally moving along a single path. The output of one step may indicate that we should take a different tack. Having rejected sameness of material substance and sameness of immaterial substance, Locke does not attempt to come up with a third sort of substance as the ground. He does not assume that enduring identity must be based on sameness of substance, and then take his project to be to figure out what sort of substance that

might be. Rather, he takes his explications to reveal that sameness of substance won't do the job he want done. Attending to why his first two proposals proved inadequate, he proposes a different criterion: Sameness of organism. Vegetables and animals, he says, retain their identity when they grow and shrink, because they are organized bodies. The matter may change over time, but it is the organization that unites an organism into a continuous living thing. Different sorts of organisms are differently organized. That is why a radish is different from a man. But a man retains his identity as a man so long as he is the same living human being, just as a radish retains its identity as a radish so long as it is the same living vegetable. This criterion does better than sameness of substance, since it blocks worries about weight gain and transmigration of souls. It respects the intuitions about the counterfactuals that sameness of immaterial substance could not accommodate. But it has consequences which might give us pause. The criterion does not track consciousness or rationality. What matters on this proposal is only the continuity of a living (that is, suitably organized) human body. It has the result, for example, that the embryo (or at least a neonate) is the same person as the adult. It also has the result that anything that is configured as a human – as Locke says, 'though it had no more reason all its life than a cat or a parrot' – would qualify as a person (p. 446). A human being in a permanently vegetative state is the same person as the active, engaged fellow he once was. Locke admits that there is some sort of identity through time in these cases, but it is not what we seek when we ask about enduring personal identity.

Again the criterion fails, but Locke has made some progress. The objection to characterizing the organism who is configured as a human but has no more reason all its life than a cat or a parrot gives us new clear foils – (imaginary) human-shaped organisms incapable reason. Locke hedges, objecting to human-shaped cats and parrots who *never in their lives* had reason. Maybe, he hints, the *loss of reason* is compatible with personal identity. Still the objection indicates that something

about reason is critical to the enduring identity of a person. He then takes up different case. Having heard that parrots can talk, he argues that with enough evidence that parrots continuously engaged in rational discourse, we would still would not consider them human, but we would consider them rational animals – that is, persons. Human, evidently, is a biological kind; person is not.

This leads to Locke's final criterion: enduring personal identity is a matter of unity of consciousness. He says, '[A]s far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person' (449). Hence if Cesar Borgia had been conscious of having done Socrates' acts, Borgia would be the same person as Socrates. If two consciousnesses inhabited the same human body, neither of whom was conscious of the actions of the other (as in the case of multiple personality disorder), there would be two distinct persons in the same body. And if a human being, perhaps because of brain damage, was utterly unable to remember the actions of his earlier life, he would no longer be the person who did those things.

Locke endorses these conclusions because, he says the concept of a person, hence the criterion of personal identity through time, is forensic. That is, it is a device for assigning reward and punishment, praise and blame. This may seem to come out of the blue. In the discussion of sameness of substance there was no mention of responsibility. But the trajectory of the argument shows how, as we refine a concept via epistemic iteration, we also refine our views about the role that concept should play. We discover first that rationality is important to personal identity, and then that it is important because it is closely tied to responsibility. Locke's final aim here, evidently, is to contour a concept of personal identity that would enable us to properly assign responsibility for past actions. The adequacy of the fruits of his epistemic iteration should be judged by reference to that aim.

We know, of course, that the iterations did not end with Locke. Every first year philosophy student can see further worries that have to be faced. Suppose Jane at age 40 recalls having adopted a dog when she was 20. Jane then is, by Locke's criterion, the same person at 40 that she was at 20. Jane at age 80 recalls having had a cat when she was 40, but remembers nothing about her actions at age 20. Hence, Jane at 80 is the same person as Jane at 40 and Jane at 40 is the same person as Jane at 20, but Jane at 80 is not the same person as Jane at 20. Problem: identity is transitive. Then there is the worry about the status of individuals with dementia who remember their childhood in detail but have no memory of what they had for breakfast. Parfit's cases about fission and fusion to complicate the issue further. The problem has not been solved.

Still, we don't go back to square one. We build on Locke. We acknowledge that his epistemic iterations did not go far enough. But we not inclined to take sameness of person to be sameness of substance (either physical or spiritual), and are at least doubtful about taking it to be sameness of living organism. We may, however, be willing to entertain the possibility that personal identity somehow supervenes on sameness of physical organism. If we want a forensic concept, we see that we have to link it to accounts of who is responsible, and where – owing to psychosis, dementia, or other incapacity or impairment – responsibility ends. If we want a more generic concept, we still have to ask how responsibility fits in the picture.

We might, for example, agree with Locke that something about a temporally extended connected capacity for rational thought is the key to enduring personal identity, but recognize that memory is not restricted to episodic memory of particular events, and that not all acquired attitudes that bear on reason and responsibility are matters of memory. Henry's excessive fear of frogs, which is due to an unfortunate, repressed event in his childhood, may still figure in his enduring identity, and in the reasons he adduces and acts and omissions that he is responsible for. We might

follow Kant and insist that a temporally extended synthetic unity of apperception is needed. There are different paths forward.

I am not especially interested in personal identity, so I won't carry this iterative sequence further. My goal has been to illustrate how, through epistemic iteration, we test clear cases against available criteria; refine the criteria by reference to the evolving sequence of classes of clear cases and the clear cases by reference to the evolving criteria. Our epistemic goals evolve in tandem with our explicative resources. Locke made progress. We understand the issues surrounding personal identity through time better as a result of his iterations.

The Problem of the Criterion

An objection might be raised to the whole iterative procedure. Chisholm characterized it this way:

To know whether things really are as they seem to be, we must have a *procedure* for distinguishing appearances that are true from appearances that are false. But to know whether our procedure is a good procedure, we have to know whether it really *succeeds* in distinguishing appearances that are true from appearances that are false. And we cannot know whether it does really succeed unless we already know which appearances are *true* and which ones are *false*. And so we are caught in a circle (1982:62).

This is the problem of the criterion. Chisholm regards both criteria and judgments about specific cases as static. The question, as he frames it, is whether we have an independent reason to credit either without relying on the other. He thinks we do not. To break the circle, he prefers privileging particular judgments rather than general procedures. He recognizes that others would favor privileging procedures instead. But, he concludes, either way, we are forced to a measure of epistemic hypocrisy. He says, 'What few philosophers have the courage to recognize is this: we

can deal with the problem only by begging the question' (1982:75). He assumes that the solution must be static and thinks that a static solution must be question-begging.

I have suggested that epistemic iteration provides a way to evade the problem without begging the question or reasoning in a circle. We are not forced to be hypocrites. Although a criterion is tested against a class of individual judgments, and the class of judgments is tested against a criterion, the criterion that tests a particular class of judgments is not itself tested *by the particular class that it itself tests*. The process is a dynamic sequence of carefully choreographed steps. Rather than privilege any class of particular judgments or any class of procedures, we use each to put pressure on and, if need be, to refine the other. We do not assume that either is fixed or final; we do assume that they are a good enough base to work from. Epistemic iteration is dynamic. Criteria and judgments about individual cases are revisable, and are revised to bring them into better accord. Our best available criteria may not be sufficient to cover every imaginable or even every actual case. But over time, we revise them and our conceptions of the cases to bring about or at least approach an alignment. Followers of Chisholm might insist that criteria *must* be fixed; a criterion is then a determinate, rigid, inviolable rule. If so, then what I have been calling criteria might be redescribed as increasingly rigorous rules of thumb. No matter. However they are characterized, they are the considerations that we use to decide whether a given concept applies in a given case. Rather than a circle, a sequence of epistemic iterations is a spiral. Perhaps that does not enable us to '*know* whether things *really* are as they appear to be' but it does give us reason to believe that our views about the correlations between appearance and reality – between concepts and conditions on their instantiation – are improving.

Nevertheless, my description of Locke's discussion makes the process look too linear. I focused, as Locke does, on the relation between concepts and criteria, and emphasized how each

can motivate the revision of the other. But epistemic iteration refines aims, norms, methods, and modes of inference in tandem. Explication figures in how it does so. At the outset of Locke's discussion, there was no hint that the concept of an enduring person had a fundamentally cognitive or fundamentally ethical dimension. By the end, it is clear that something about rationality and responsibility should figure in an adequate concept. Moreover, we now take it that an adequate concept must accommodate intuitively correct verdicts about counterfactual cases, even ones that we are confident will never obtain. We recognize that the philosophical acceptability of a concept depends not just on the way the world is, but also on the role the concept is to play in our cognitive economy. As we refine the end, we refine the requirements on what it takes to achieve that end.

We do the same process when testing aims against criteria for their satisfaction; methods against the criteria of success; modes of inference against criteria of validity; and so forth. Nor do we do these one by one. We ask whether our aims with their criteria of satisfaction can be realized given the available methods with their criteria of success, as judged by the methods with their criteria of validity. We are willing to modify any in light of the others to improve our overall, evolving epistemic aspirations. Moreover, other fields of inquiry do the same. They fine-tune their concepts, methods, standards, and modes of inference against one another in order to promote their evolving agendas. And, however parochial their specific questions and methods are, they recognize that they must respect requirements like consistency and coherence that apply generally, and must align with the commitments of adjacent disciplines. If this is right, then as we improve our concepts, methods, and standards, we refine and correct what we earlier thought, and we expand our understanding of how things in the broadest possible sense of the term hang together in the broadest possible sense of the term.³

³ I am grateful to Arham Aftab Kazi and Georg Brun for advice on an earlier draft of this paper.

References

- Brun, Georg (2016). 'Explication as a Method of Conceptual Re-engineering' *Erkenntnis* 81: 1211-1241.
- Carnap, Rudolph (1950). *Logical Foundations of Probability*. London: Routledge and Kegan Paul.
- Chang, Hasok (2004). *Inventing Temperature*. Oxford: Oxford University Press.
- Chisholm, Roderick (1982). 'The Problem of the Criterion' *The Foundations of Knowing*. Minneapolis: University of Minnesota Press, pp. 61-75.
- Dewey, John (1916). *Democracy and Education*.
- Elgin, S. Z. (2020). 'On Question-Begging and Analytic Content' *Synthese* 197:1149-1163.
- Locke, John (1689/1894). 'Of Identity and Diversity', *Essays Concerning Human Understanding Book II*. Oxford: Clarendon Press, pp. 239-270.
- Quine, W.V. (1969). 'Epistemology Naturalized' *Ontological Relativity and Other Essays*. New York: Columbia University Press, pp. 69-90.
- Ross, W. D. (2002). *The Right and the Good*. Oxford: Oxford University Press.
- Sellars, Wilfrid (1968). 'Philosophy and the Scientific Image of Man' in his *Science, Perception, and Reality*. London: Routledge and Kegan Paul, pp. 1-40.
- Wittgenstein, Ludwig (1961). *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.