

The Sustainability Code*

A Policy Model for Learning and Knowledge Processing in Human and AI Systems

Joseph M. Firestone, PhD and Mark W. McElroy, PhD
Co-Founding Principals, Artificial Epistemics, LLC

Introduction

Many years ago, the well-known science fiction writer, Isaac Asimov, wrote of futuristic societies in which intelligent robots would work in the service of humans (Asimov 1942). In his original formulation, Asimov's robots were indelibly programmed with what he called the Three Laws of Robotics: (1) A robot may not injure a human being, or, through inaction, allow a human being to come to harm; (2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law; and (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

When viewed from a Knowledge Management perspective, Asimov's laws correspond to an instrumental level of behavior on the part of robots. In other words, his laws apply to the actions or potential actions of robots, to things they might do in the material world. But there is a precursor to action that Asimov's laws did not address – his laws did not address learning or knowledge processing. In Philosophy, we call the study of laws of learning, or innovation, Epistemology (Audi 1998, Kirkham 2001). Let us imagine for a moment, then, what it might mean to design robots with hard-coded epistemologies, and how their capacity to learn – as well as *what* they learn – might differ, depending upon the epistemologies we give them.

Consider two identical robots, the same in every way except for their respective epistemologies – one with a Realist epistemology and the other with a Relativist one. Chances are that when faced with the same problems, these two robots will develop very different conclusions about how best to answer or act in response. Among the potential outcomes, then, will be that one or both of them will get things wrong, despite the fact that their underlying epistemologies will be telling them otherwise. Human well-being will hang in the balance. The point is that Asimov's Laws are laws of action; epistemologies, by contrast, are laws of learning and knowledge processing, upon which actions depend. Actions, that is, are knowledge in use!

Now let us turn to the real world of humans and artificial intelligence (AI) in modern times. We, too, and the AI systems we create are constrained by laws of action at many different levels of analysis. We are constrained by moral laws in our families and communities; by formal laws in our legal systems; and by administrative or procedural laws in our organizations. Our *learning* laws, or rules, however, are not so well defined. Still, we rely on them, utterly, for our survival. They determine how we think, how we view the world, and how we reach conclusions about truth and morality. So, too, it is for AI. What's required, then, are self-regulated learning systems in AI that can help guard against misinformation and unethical action.

Because action reduces to knowledge in use, the quality of our conclusions from learning matters greatly in terms of the quality of our actions, and whether or not our actions are effective, beneficial, and sustainable. Can we not say, then, that epistemology is a variable in human and artificial intelligence? We believe we can (McElroy 2003). To the extent that action taken on the basis of truth is more likely to be effective, epistemology is very much a factor in the sustainability of our actions and outcomes. Truth and sustainability are joined at the hip – and so it is for morality as well. How best to obtain them, though, is not so clear.

Indeed, when the day comes when scientists are faced with the question of how best to endow artificial systems with a capacity to learn (i.e., *now*), which laws or rules of learning will they choose? Will the robots and electronic brains of the future be Relativists or Realists? Will they rely on the correspondence theory of truth or the coherence theory? Or will they be pragmatists or instrumentalists? And if the choices scientists ultimately make on these matters have impact on the actions taken by AI systems and their consequences, can we not say that the design and implementation of such systems must include explicit treatment of these things? And if so, where is the epistemological layer of AI, and what is the best way to design it?

*Adapted from a paper by the same title presented by Mark W. McElroy, PhD at the 4th International Symposium on Soft Science, Beijing, China, November 28-30, 2006; and as published again in 2016 in *The MultiCapital Scorecard* by Thomas and McElroy.

The Sustainability Code

In 2005, the two of us sat down to expressly take up the questions above – not only about artificial systems, but human beings, too. We asked ourselves: *If what people on earth want is a pattern of action for both human and AI systems that is effective, beneficial, and sustainable, what must their learning systems be in order to meet those goals? If reliable learning is the precursor to safe, ethical and effective action, what rules, laws, or principles of learning should we have in order to maximize the quality of our learning and knowledge processing – to achieve sustainable innovation, as it were? What should our epistemology be?*

We started by acknowledging that *truth matters* in the conduct of human affairs, and that if what people want is the ability to take effective action, they should predicate their behaviors on the basis of truth, not falsity – for falsity arguably leads to *ineffective* action. Thus, it is always better to take action based on the way things *really are*. Here, we intentionally adopted a Realist epistemology (Audi 1998, Kirkham 2001), a metaphysical assumption that (1) the world does *in fact exist*, and (2) that we do in fact interact with it and can describe and evaluate it in meaningful ways.

We further took care to extend the application of our thinking to not only the truth (or falsity) of descriptive knowledge (i.e., relative to factual claims), but also to the legitimacy (or illegitimacy) of normative or evaluative knowledge (i.e., relative to norms and value claims). Knowledge, that is, is by no means limited to descriptions of facts in the world, but also includes knowledge that is evaluative of such facts or which separately expresses claims about the way the world ought to be and how humans (and AI) ought to behave. Of greatest value, then, would be an epistemology that is equally applicable to both facts and values!

Next in our formulation was a decision to give priority to sustainability as a desired outcome from effective learning – that learning should make *subsistence and adaptation* possible, and in no way undermine or conflict with human social, economic or environmental well-being. We thereby called our set of rules, then and now, the *Sustainability Code* – 'Sustainability' because we saw learning as an adaptive strategy for both living/human and artificial systems, and 'Code' because of the prescriptive or regulatory sense in which that word is often used. What we were trying to create, then, was a prescriptive model for use at the epistemological level of human and AI systems – a specification for safe, reliable, and sustainable knowledge processing; systems designed to tell us the truth and keep us safe and free from the dangers of misinformation and immorality (see Figure 1).

Thus, the *Sustainability Code* (aka, the *Susty Code*) was born – a policy model for human and AI systems that designers and managers can use in their attempts to cultivate sustainable innovation, a pattern of knowledge processing and problem solving that is safe and reliable. As such, the *Susty Code* is a target-state model, a target state that can be used as a 'blueprint' or specification for both human and artificial epistemologies, with an eye towards achieving sustainability in the conduct of human affairs.

Eleven Rules

Let us now examine the components of the *Susty Code*, a prescriptive set of eleven rules, or policies, for learning and knowledge processing:

1. All knowledge used as a basis for individual and/or collective action shall always be open to criticism, and no such knowledge shall ever be regarded by any member or constituent as true or legitimate with certainty. This is the **FALLIBILITY** rule. It applies to both human and AI systems.

This rule stems from (a) a Realist epistemology, and (b) the conviction that all human and AI knowledge is irreparably fallible (Popper 1972, Notturmo 2001). Thus, it would be risky and unsustainable to take action on the basis of the view that knowledge of any kind is correct with certainty, since it would only serve to insulate potentially false or illegitimate knowledge from criticism or correction, and thereby expose humans to undue risks arising from actions predicated on mistakes.

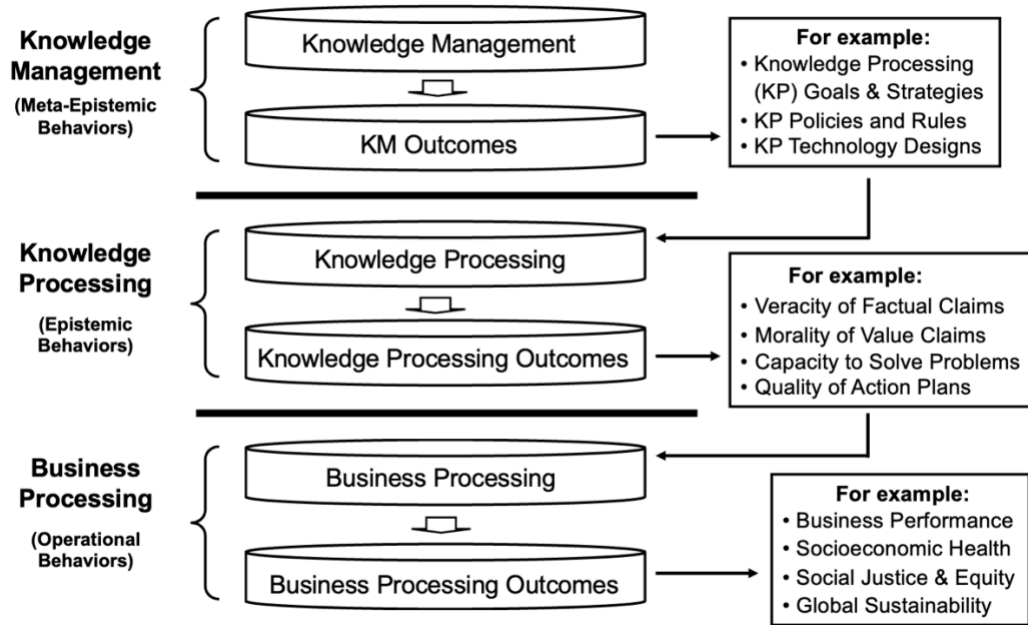


Figure 1 – The 3-Tier Model of Knowledge and Business Processing

2. All information and knowledge used by individuals or collectives shall be accessible and transparent to all human and AI collectives, regardless of management roles or structures in place. No such information and knowledge shall be withheld from individuals or members of the collective by another, except in cases where fulfilling fiduciary duties, including the protection of trade secrets, or the need to respect privacy entitlements are involved. This is the **TRANSPARENCY** rule.

The principle of transparency is fundamental to effective learning, innovation, and survival. For how can we expect anyone to adapt to circumstances in the world if information about them is hidden from view? Indeed, opacity, the opposite of transparency, is risky and unsustainable as a policy for learning, adaptation, and innovation.

3. All learning and knowledge processing in human and AI systems shall be accessible to, and inclusive of, all members, regardless of whatever separate and/or restricted management roles or structures may be in place. This is the **INCLUSIVENESS** rule.

We sometimes refer to this concept as *epistemic inclusiveness*.¹ What it means is that stakeholders, or members/agents of a human or AI system, must be permitted to have access to, and participate in, the learning, innovation, and knowledge processing activities of the collective. Excluding individuals from such processes only engenders resentment, and deprives the broader population of its own members' capacities to learn, solve problems, and adapt.

4. All members of human and AI systems shall employ their best efforts to seek, recognize, and formulate problems in existing knowledge through critical evaluation of the performance of such knowledge in action. This is the **LOOKING FOR TROUBLE** rule.

The purpose of learning and knowledge processing is to help us adapt by allowing us to close our epistemic, or knowledge, gaps. Learning is our adaptive strategy (Holland 1995). Thus, the most adaptive human and AI systems will be those in which the search for epistemic gaps is a deliberate and continuous one (see Figure 2). Trouble, in this regard, will consist of epistemic gaps that we might not be aware of, but which could be discovered if only we looked for them.

¹*Epistemic*: An adjective defined as “of or having to do with knowledge or the act or ways of knowing” (*Websters Dictionary*, 2016).

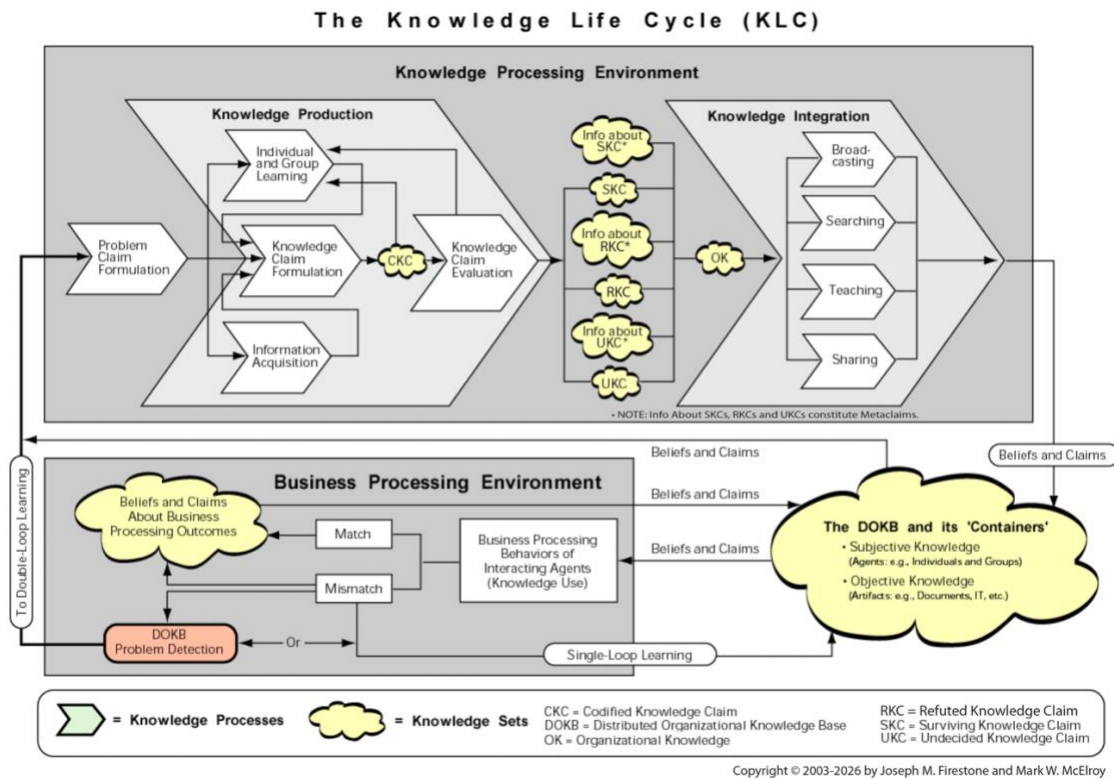


Figure 2 – The Knowledge Lifecycle (KLC)

5. All members of human or AI systems may produce any new rule not otherwise specified by these rules, so long as it and the knowledge processing system used to produce it do not contravene them. This is the **GROWTH OF KNOWLEDGE** rule.

Human well-being, sustainability, and effective action require the continuous production of new knowledge in order to make adaptation possible (Holland 1995). The fact that we have established these learning-related rules is not to say that knowledge of other kinds cannot, or should not, be produced. Or that the rules themselves are not imperfect.

6. All learning and knowledge processing in human and AI systems shall be rooted in the principle of *fair critical comparison*, such that prevailing or competing knowledge claims may always be criticized, tested, and evaluated against one another in a fair and systematic way. This rule shall also apply to claims about what such tests themselves should consist of, and not just to the primary claims to which they may be applied. This is the **FAIR COMPARISON** rule.

This principle stems from the distinction we can make between theories of truth/value and theories of evaluation. Even when we have settled on theories of truth or value, such as the Realist theory and the correspondence method that usually accompanies it, we are still left with questions about how best to test and evaluate competing beliefs or claims. The *Fair Comparison* test is a specific theory of evaluation originally developed by one of us in 1974 (Firestone 1974). It was later incorporated into a theory of Knowledge Management put forward by Firestone and McElroy, known as *The New Knowledge Management* (Firestone and McElroy 2003, McElroy 2003), and significantly enhanced thereafter. It is arguably the centerpiece of the *Sustainability Code* in terms of the role it plays in making *self-regulated* knowledge production possible, and the proportion of the *Code* it accounts for (roughly 80 percent).

7. Next is the all-important distinction we must make between our knowledge of facts and our knowledge of values. Broadly speaking, factual knowledge is descriptive in content and consists of beliefs and claims about the world in terms of the way it is, the way it has been, and the way it will be. Values

knowledge, by contrast, is evaluative and/or normative in content and consists of knowledge about the world in terms of how good or bad it might be and also what it ought to be. Knowledge about duties, for example, is values knowledge.

Here it is important to understand that both kinds of knowledge are subject to the same principles of learning and knowledge processing, and that survival and sustainability in the human condition is not just a function of our knowledge of facts, but our knowledge of values as well (Hall 1961). The distinction between the two and their relevance to human well-being must therefore be acknowledged in all human and AI systems as core components of their knowledge and information ontologies. This is the **FACT/VALUE** rule.

Understanding and assessing the way the world is and ought to be is essential to human well-being and to our capacity to learn and adapt. It is vitally important, therefore, that human and human/AI systems be well versed and fluent in the distinctions we can and must make between facts and values. How else can we effectively comprehend and function in the world around us?

All that said, we must be clear that we are not proponents of the fact-value dichotomy as it has been applied to individual statements or claims. We do not believe in a radical logical gulf between "is" and "ought" statements. Instead, we think that *networks of knowledge claims have both descriptive and evaluative aspects* and that the essence of objectivity is to test each of these aspects according to the criteria of fair critical comparison (Rule 6).

8. All designers and managers of human and AI systems shall establish Knowledge Management functions that will be independent of the Executive Function and vested with enforceable authority to (1) allocate resources for enhancing leaning and knowledge processing in their environments, (2) manage, change, and enhance knowledge processing rules as needed, (3) handle crises in knowledge processing, and (4) negotiate for resources with other system-wide functions. This is the **KNOWLEDGE MANAGEMENT** rule.

Rules and policies for learning and knowledge processing require management, if only to aid in their implementation and use. This is a definition of Knowledge Management that sees itself as a management discipline that seeks to enhance the quality of knowledge processing in human and AI systems (Firestone and McElroy 2003, 2005). Moreover, it is a management discipline that any human or AI system must have if its patterns of learning and knowledge processing are to be sustainable, and themselves adaptive.

9. The actual or potential performance of knowledge in action shall be defined to include the social, economic, and environmental consequences of actions taken, and in particular the sustainability of such impacts. No such impacts shall arbitrarily be externalized or otherwise excluded from the scope of evaluations performed under rule number 6 above (Fair Comparison), and all such impacts determined to be unsustainable shall be internally costed accordingly in related evaluations. This is the **INTERNALIZATION** rule.

This is a very important rule that goes to the very heart of the sustainability crisis in the world today (Daly 1996). To the extent that human systems in particular (i.e., businesses) are externalizing many of their negative social, economic, and environmental costs, knowledge of such costs as a precursor to taking actions is also being externalized. This is a prescription for disaster in the conduct of human affairs, and it must be rejected if we are to make any progress in improving the sustainability of our behaviors. Thus, knowledge of externalized social, economic, and environmental impacts (and their costs) must be internalized in the learning routines of a collective, as if the impacts (and costs) were internal to the collective itself. The performance measurement and reporting systems of businesses in particular, therefore, should be structured, accordingly (McElroy 2008, Thomas and McElroy 2016).

10. The Knowledge Management function shall adopt and implement only knowledge processing policies that are aligned or synchronized with the self-organizing tendencies of human collectives to produce and integrate knowledge in their own ways (again, see Figures 1 and 2). This is the **POLICY SYNCHRONIZATION** rule.

People have a tendency to self-organize around the discovery of problems (epistemic gaps) and the knowledge production and integration activities that follow (Figure 2). Knowledge production in human and AI systems is therefore an emergent process (Stacey 1996). Strictly speaking, then, no manipulation or management is required to get people to learn or innovate; human social systems, in particular, are already endowed with predispositions to do so. Thus, a sustainable human or AI system will be one that welcomes and supports these tendencies, and which does not either (a) conflict with or undermine them, or (b) engender learning or innovation outcomes that have the effect of working *against* people and not *for* them (McElroy 2003). Indeed, a sustainable innovation system will be one that actually helps people to adapt, not mal-adapt. This in turn calls for the implementation of policies which are aligned, or *synchronized*, with human tendencies to self-organize around the production and integration of new knowledge in their own characteristic ways.

11. Any member of a human or AI system who fails to abide by these rules shall be subject to exclusion from the collective by its other members at their discretion. This is the **ENFORCEMENT** rule.

Collective living need not amount to a suicide pact in the event that some members choose to put others at risk by failing to abide by reason, rationality or ethical integrity. To the extent that the *Susty Code* is a policy model intended to enhance the capacity of individuals and collectives to learn, adapt, survive, and live in sustainable ways, members who violate these principles may justifiably be excluded from the community by their peers who remain committed to sustainability, truth, and respect for human rights.

Conclusions

For an organization or society to be sustainable, it must have two things: (1) knowledge of its impact in the world, and (2) the ability to learn or innovate in response. For this reason, conscious attention must be paid to managing the epistemologies of human and AI systems, for it is the epistemology of a system that makes learning, adaptation, and effective action possible. Thus, we can say that any epistemology or knowledge processing system that meets these two criteria is sustainable, and any that does not, *is* not.

The concept of sustainable learning and innovation, then, rests in part on the distinction between learning and action. In other words, we can differentiate between (a) the sustainability of what a business, for example, does, and (b) the sustainability of the internal knowledge processes it relies on for problem solving and learning. Moreover, we can say that unsustainable knowledge processes will more often beget unsustainable behaviors and outcomes, and that such knowledge processes are more likely to work against us than for us.

It should also be clear that in many contemporary human social systems, a majority of the rules set forth above are missing. While the predominant corporate epistemologies in most multi-national firms, for example, are admittedly Realist in form, their learning-related policies are too often regressive and ill-defined. In most such cases, organizational knowledge is *justified* by appealing to the authority of management (violates Rules 1, 4, and 6 above), with most *official* organizational knowledge being too closely held and cultivated by management alone (violates Rules 2, 3, 5 and 10 above).

Of even more concern is the failure of most organizations to fully take their external social, economic, and environmental impacts into account as they make plans and assess their own performance (violates Rules 7 and 9). In the process, not only are the full factual implications of organizational operations overlooked, but so are the evaluational or value-based reactions we can have when confronted with our impacts in the world around us.

Alas, most organizations take a rather reductionist and mechanistic approach to the management of learning and innovation, choosing to solve artificial problems with artificial processes, and thereby miss out on benefiting from the real potential of human creativity and problem solving (violates Rule 10 above). Instead, aberrations of our Rule number 11 are enforced, according to which people are punished or excluded from a collective simply because of their desire to participate in the innovation processes of the organization.

And then there is the AI domain in its entirety, which so far either ignores the epistemological layer of a system per se, or relies instead on philosophies of knowledge processing that, while common, are notoriously problematic (e.g., fallacious appeals to authority as a basis for truth) (Cavender and Kahane 2010). Indeed, what we see being used in most cases are AI systems in which the epistemological policies in force are either implicit at best, haphazardly configured, or both; not to mention negligent in terms of the attention they give to *value* claims as compared to purely descriptive ones. Ethics and morality suffer as a consequence.

In sum, the key questions addressed by this paper are: *Do designers and managers of human and AI systems give sufficient attention to epistemological factors within their environments? And are their knowledge processing systems safe, rigorous, and reliable?* Our answers to both questions are *No*, and that *Designing AI systems, in particular, without taking such factors explicitly into account, is reckless and irresponsible.* Why? Because faulty (if not missing) knowledge processing systems too often leads to bad actions and outcomes. To guard against these things, developers must carefully design knowledge processing functions in accordance with the kinds of Rules we offer above, but which must be specified in much greater detail.

The theory presented here, then, is that learning and knowledge processing is a variable process, the makeup of which can either help us or hurt us, depending on whether it facilitates reliable understandings of the world around us and norms for how best to answer or act in response. To the extent that it does those things well, it can serve our purposes; to the extent that it doesn't, it can be our undoing.

References

- Asimov, I. (1942) "Runaround." *Astounding Science Fiction*, March: 100.
- Audi, R. (1998) *Epistemology – A Contemporary Introduction to the Theory of Knowledge*. London: Routledge.
- Cavender, N. and Howard Kahane (2010) *Logic and Contemporary Rhetoric – The Use of Reason in Everyday Life*. Belmont, CA: Wadsworth.
- Daly, H. (1996) *Beyond Growth*. Boston, MA: Beacon Press.
- Firestone, J. (1974) *The Adaptive Crisis and the Foundations of Social Science: A Critique of Empirical Social Science and Some Suggestions for its Reconstruction* (unpublished manuscript). Binghamton, NY: State University of New York at Binghamton.
- Firestone, J. and Mark W. McElroy (2003) *Key Issues in the New Knowledge Management*. Burlington, MA: Butterworth-Heinemann.
- Firestone, J. and Mark W. McElroy (2005) "Doing knowledge management." *The Learning Organization Journal* (12)2.
- Hall, E. (1961) *Our Knowledge of Fact and Value*. Chapel Hill, North Carolina: University of North Carolina Press.
- Holland, J. (1995) *Hidden Order – How Adaptation Builds Complexity*. Reading, MA: Perseus Books.
- Kirkham, R. (2001) *Theories of Truth – A Critical Introduction*. Cambridge, MA: MIT Press.
- McElroy, M. (2003) *The New Knowledge Management – Complexity, Learning, and Sustainable Innovation*. Burlington, MA: Butterworth-Heinemann.
- McElroy, M. (2008) *Social Footprints – Measuring the Social Sustainability Performance of Organizations* (a published dissertation). Groningen, Netherlands: University of Groningen.
- Notturmo, M. (2001) *Science and the Open Society – The Future of Karl Popper's Philosophy*. Budapest, Hungary: CEU Press.
- Popper, K. (1972) *Objective Knowledge – An Evolutionary Approach*. Oxford, UK: Oxford University Press.
- Stacey, R. (1996) *Complexity and Creativity in Organizations*. San Francisco: Berrett-Koehler.
- Thomas, M. and Mark W. McElroy (2016) *The MultiCapital Scorecard – Rethinking Organizational Performance*. White River Junction, VT: Chelsea Green Publishing.

For more information about the *Susty Code*, contact Mark W. McElroy at Artificial Epistemics: mwm@sustycode.ai. All related disclosures are made at the sole discretion of Artificial Epistemics, LLC, subject to the terms of Non-Disclosure Agreements (NDAs).