



# NEUROETHICS, THEORETICAL GUARANTEES, AND SOCIOTECHNICAL RISK

MARK FEDYK

# PRELUDE



# Question

What is “the ethics” of an NSF sponsored **Science and Technology Center** organized to push brain-machine interface technology forward to answering grand questions and achieving grand applications?

**Science and Technology Centers (STCs):  
Integrative Partnerships**

Special STC News Announcement



The U.S. National Science Foundation announced six new Science and Technology Centers to advance ambitious, complex research in fields ranging from mechanobiology to particle physics to climate change. For decades, NSF Science and Technology Centers have transformed cellular biology, combined scientific disciplines to enhance accelerator capabilities, and revolutionized real-time functional imaging by providing the ability to observe the activity of a single atom.

# "The Ethics" of Anything

- Well, what is "the ethics" of *anything*?
- It is **a conceptual framework** that
  - Defines some good, some bad, some right, and some wrong.
  - Describes the causal origins and/or logical foundations of things that fall into those categories.
  - Explains how new ethical knowledge about new g/b/w/r can be acquired.
- If that is the **general form** of "the ethics" of some X, then...
- "The ethics" of the BMI STC is then a **special case** of this.
- Details matter immensely, however.

# First: Thank you!

- Professor **Karen Moxon** has been **extremely** generous with her time, going around a few circles – both literally and logically – propelled by Socratic questions in order to figure out the parameters of the **special case**.
- Thanks also to Professor Étienne Brown and Professor Tina Panontin for their time and discussion about this project.
- And thanks also to the attendees of the Division of General Medicine, Geriatrics, and Bioethics November **RFLIP** session, who gave me helpful feedback on a previous version of this talk.



**UCDAVIS**  
**HEALTH**

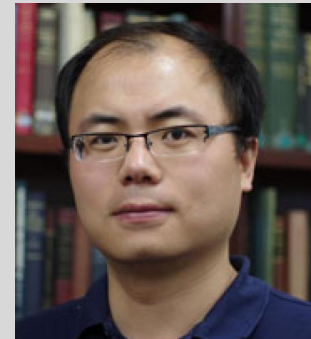
**SCHOOL OF  
MEDICINE**

Department of Internal Medicine

Division of General Medicine, Geriatrics, and Bioethics

# First: Thank you!

- Thanks as well to Dr. Alok Srivastava, Dr. Gautam Kumar, and Dr. Zhaodan Kong for permission to use some of their ideas and slides – and more importantly for passing along some important ideas and maps.



# Building out the Special Case

- What is a STC?
- It is (what philosophers and sociologists of science call) a **research programme**.
- It is a small- to medium-sized network of scientific collaborators who share core scientific commitments -- "**they speak the same scientific language**" and "**know how to play together with the same scientific toys**" --, and
- Who must design and build any number of processes - both human and non-human - in order to achieve both planned and unplanned scientific ends. Call these processes **workflows**.(\*)
- Longest-term scientific ends - for example: near 100% bi-directional fluency between cortically-represented decisions and BMI devices - are more stable, while medium-term and short-term ends emerge and must be brought into equilibrium with existing workflows and exogenous factors like grant-making politics.
- **Corollary**: "the ethics" of the STC is going to be a "workflow among workflows", with its own short-term and medium-term ends - ends which should contribute (and therefore be tweaked and recalibrated so that they contribute) to the stable long-term ends.

(\*) cf. Clarke & Gearson 2009

# Today's Talk

- Some proposals about how the "workflow among the workflows" - "the ethics of the STC" - might be designed.
- Recall the definition of "ethics" above - this involves constructing with a framework useful for defining short-term and medium-term goods, bads, rights, and wrongs (g/b/r/w).
- So, we need to look at some of the short-term and (early) medium-term g/b/r/w.
- Long term g/b/r/w might be too hard to forecast, but it doesn't mean we can't put some thought into processes (=s elements of workflows) that prepare us to address these when the (in the future) we transition from (today's) short-term to (today's) medium term.
- Goals are therefore twofold:
  - #1 **Describe the conceptual framework**, and
  - #2 **Sketch some of the practical details** as to how this framework can serve as the foundation of a "**workflow among the workflows**".

The background of the slide is a vibrant, abstract composition. The top half features a solid orange background. Below this, a series of overlapping, semi-transparent geometric shapes—primarily triangles and circles—create a dynamic, layered effect. These shapes are colored in a spectrum including red, purple, blue, green, and yellow. The bottom half of the slide is a solid, light gray. The text "PART 1: STAGES OF RESEARCH" is centered in the gray area.

# PART 1: STAGES OF RESEARCH

# How do you tell the difference between “short-term” and “medium-term” and “long-term”?

- If “the ethics” is going to be a “**workflow among the workflows**” then it must be **stage of research relative**.
- If “the ethics” is going to be **stage of research relative**, then we need a model that can tell us where in the research process we are located.

# A Model of Stages of Bioengineering Research

Here I'm going to rely on a model built by a friend and colleague, Dr. Alok Srivastava (*Biological Engineering Collaboratory / Playful Dyads*).

- PhD in Molecular Biology & Biophysics (MIT 2000).
- After working at a few biotech companies, began cross-training in history, philosophy of science, institutional economics, and sociology of science to try to figure out why "translation" from lab to product so frequently fails.
- Model I am about to share was first introduced in a talk titled: "**Making Wholes to Understand Them - The Application of Chemistry's 'Total Synthesis' Practices to Genes & Genomes.**" presented at the Workshop on History of DNA Synthesis and the Organism, 2017 organized by Dominic Berry & Jane Calbert of the Engineering Life Project of Edinburgh University and hosted at the Chemical Heritage Foundation in Philadelphia.
- Model I'm about to share has been validated by an ongoing ethnographic study of Bioengineering Research Labs, specifically Dr. Zev Bryant's Lab in the Stanford Bioengineering Department, doing biomolecular engineering of motor proteins and enzymes with Single Molecule Biophysics methods and techniques.

# The Idea Behind the Model

Scientists who want to engineer nature must reconcile theories and technical capacities from several different scientific origin points and traditions.

Inductive evidence that this process goes through three stages – each of which can be many decades in length:

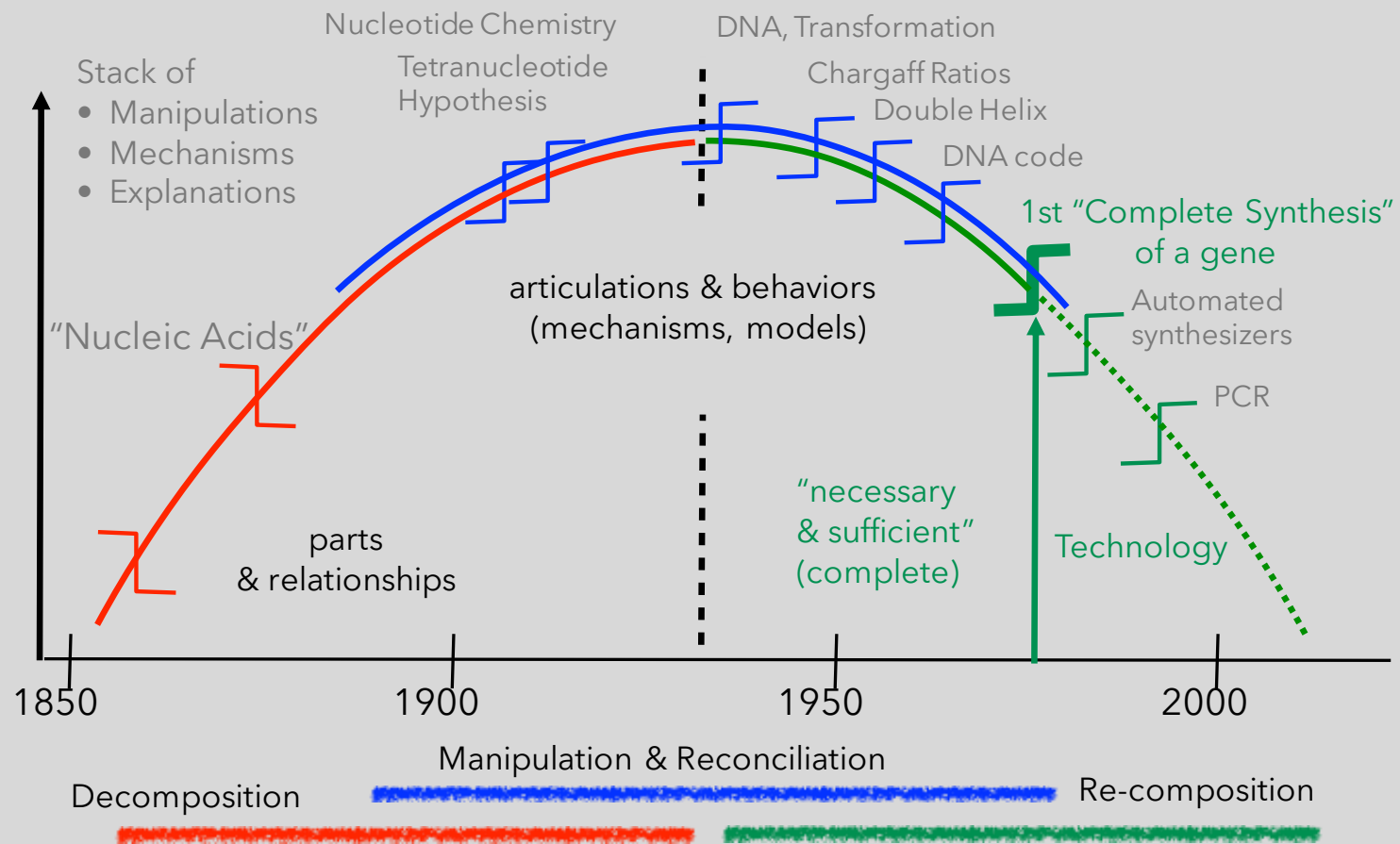
1. **Decomposition - describe “atomic” parts of natural systems**
2. **Manipulation and reconciliation - construct toy mechanisms and models that “imitate” aspects of natural systems**
3. **Re-composition - design and build useful technology, the functions of which mirror those of the natural system**

Key intuition: The architecture of the technology that “recreate” whatever natural system was successfully decomposed is (almost of practical necessity) very different than the “architecture” of the natural system – because of that knowing the components of the natural system doesn’t tell you how to build (“synthesize”) the thing that allows you to “recreate” nature.

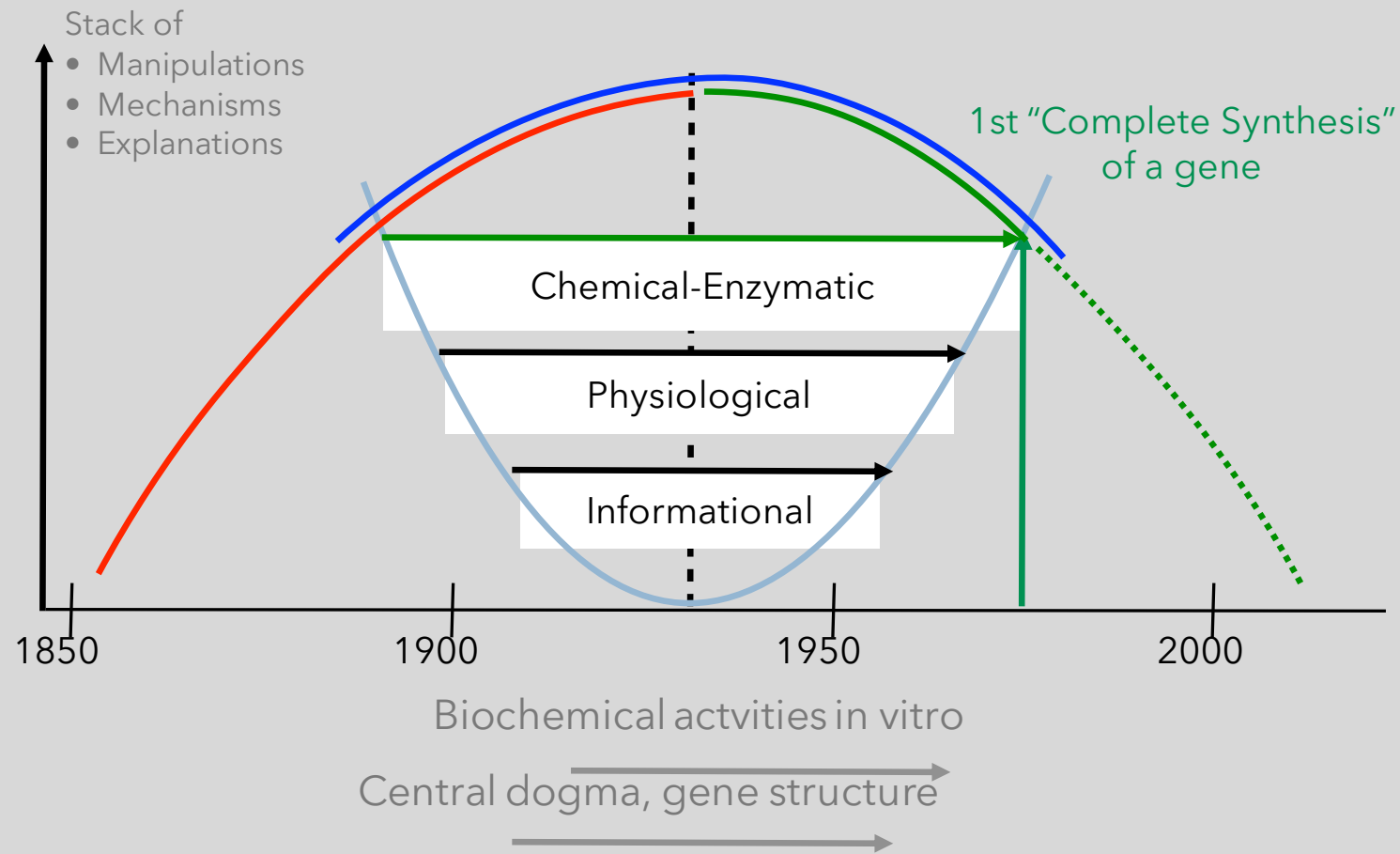
E.g. Cortical signals are constructed - generated - propagated by very different “devices” than BMI-device components.

E.g. “Executive processing & control” will in a BMI device have a different algorithmic architecture than natural cognition.

# History of DNA Science Leading to Gene and Genome Synthesis



# Moving from decomposition to recomposition is cumulative & inter-dependent



Stack of

- Manipulations
- Mechanisms
- Explanations

Reconciliation and Manipulation

BMI Device "Alpha"

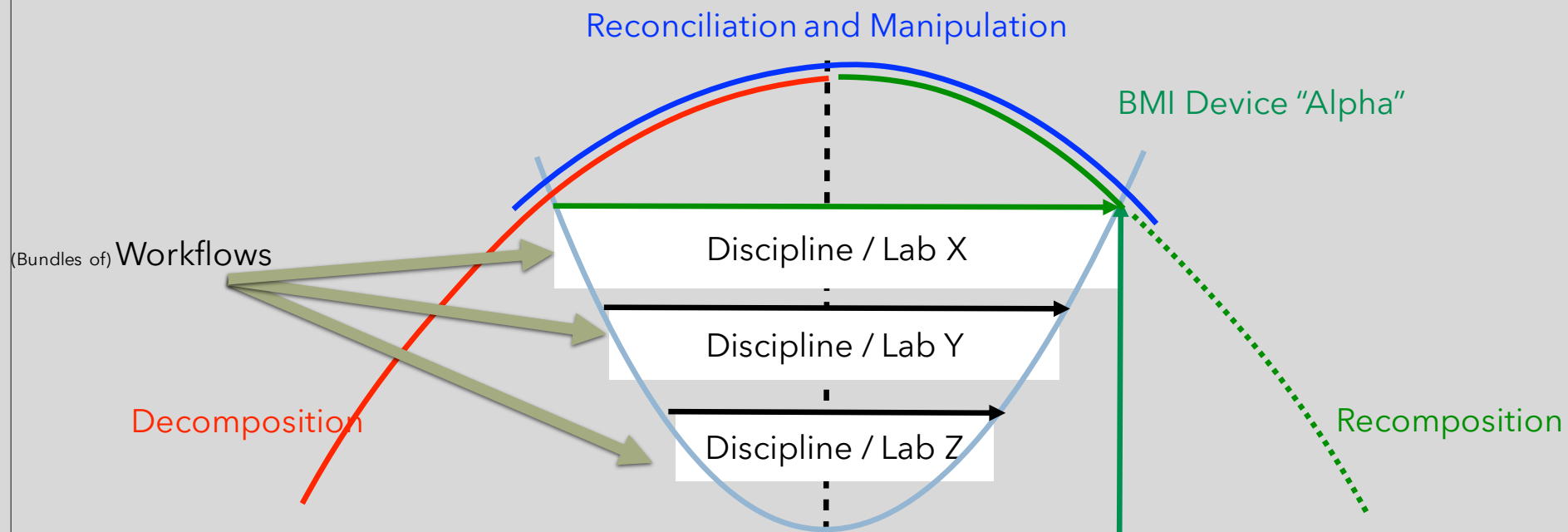
Decomposition

Recomposition

Discipline / Lab X

Discipline / Lab Y

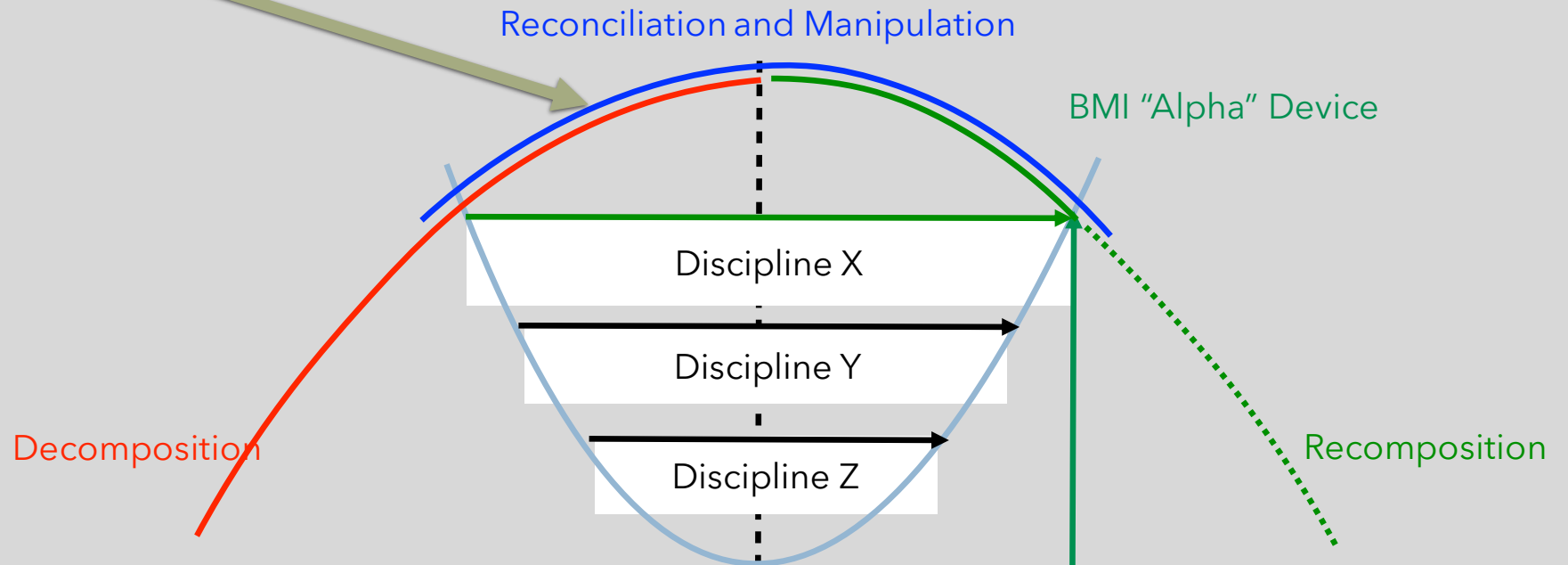
Discipline / Lab Z



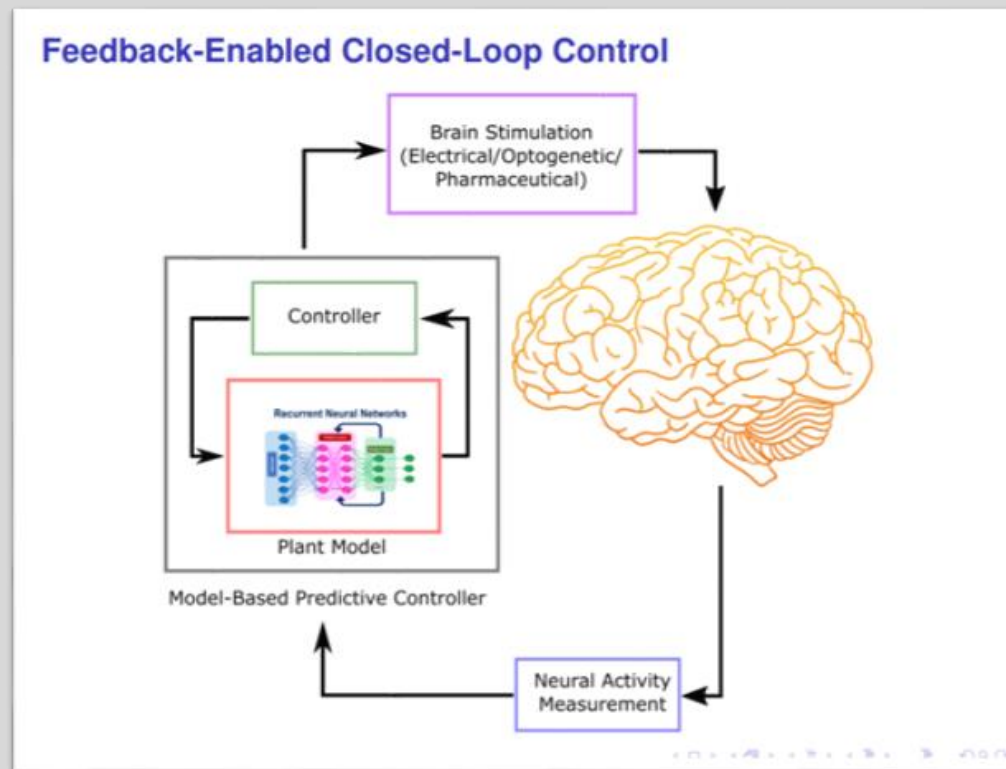
# Where are we?

- I don't know!
- But my intuition is that the science is somewhere in the middle of the reconciliation and manipulation stage, and the excitement is that people are starting to have clear ideas about how re-composition projects might succeed and dynamically mesh in unanticipated ways that prepare the ground for a period of compounding technical capacities.

Maybe we are here?

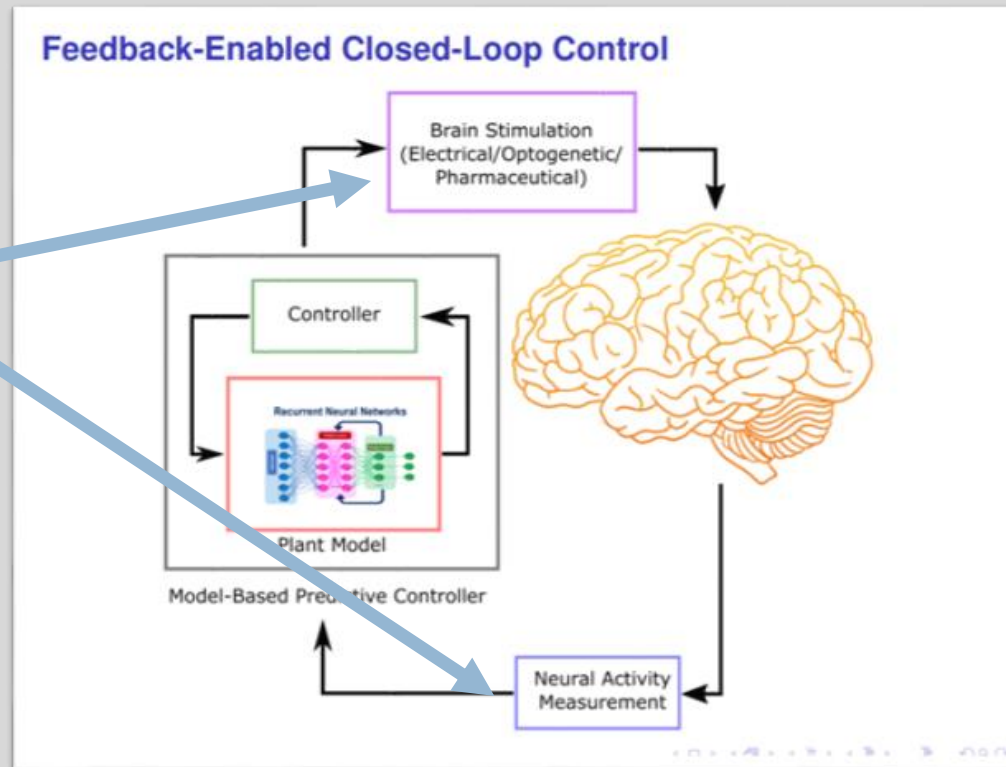


# Where are we?



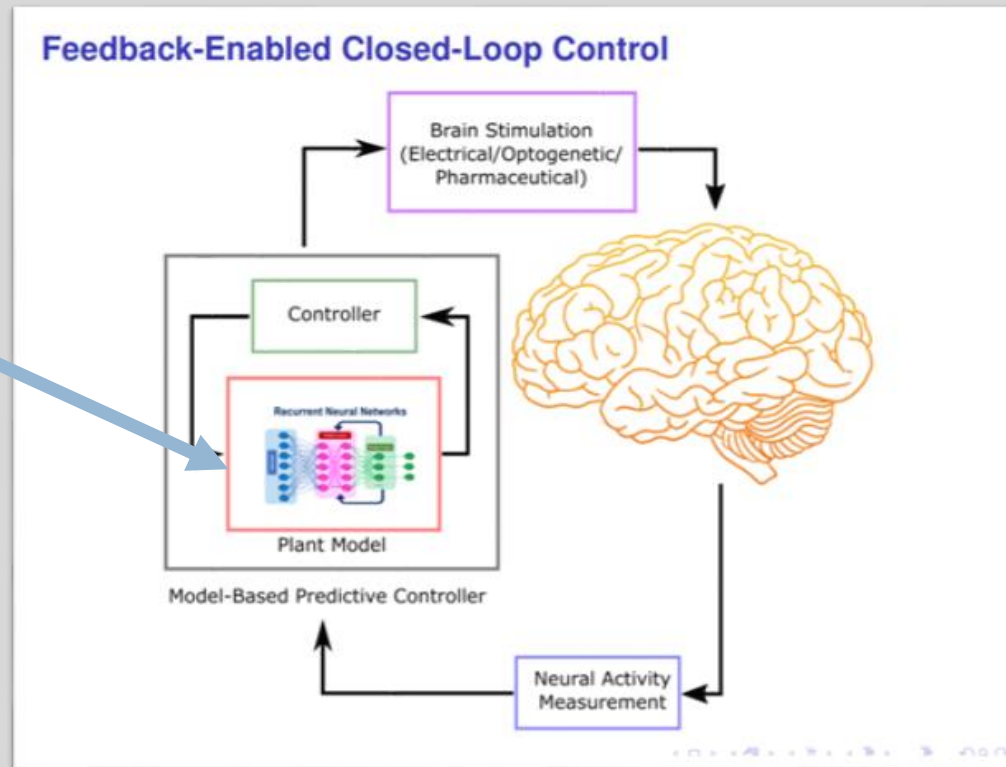
# Where are we?

Decomposition  
work here  
largely finished



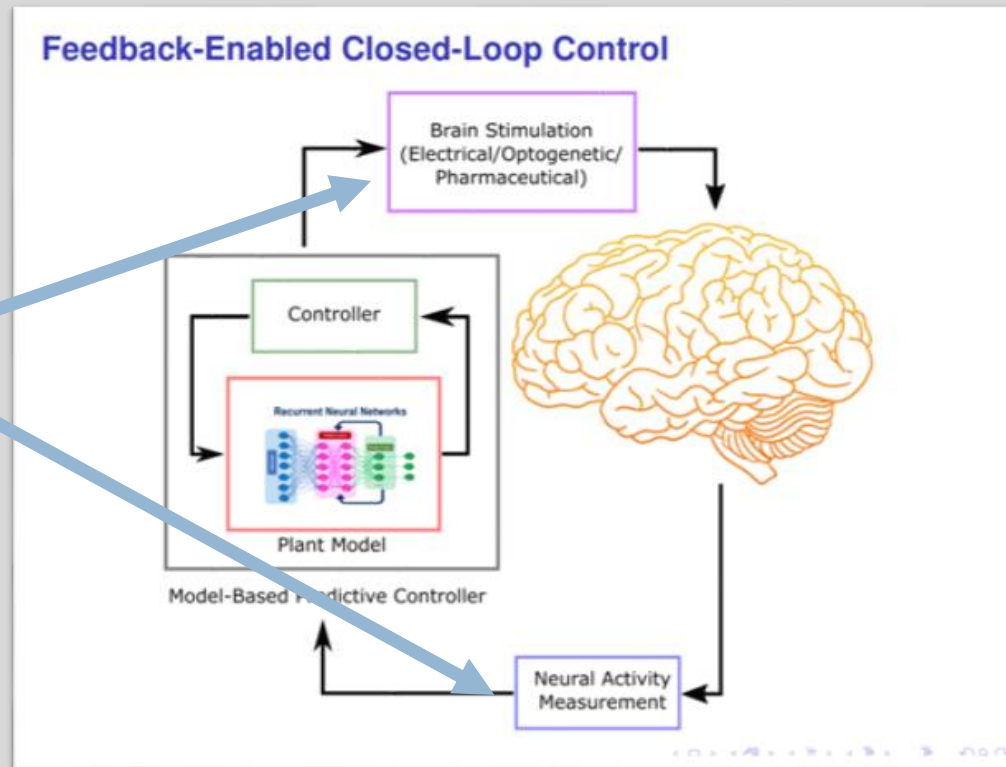
# Where are we?

A lot of the  
action seems to  
be design work  
in this segment



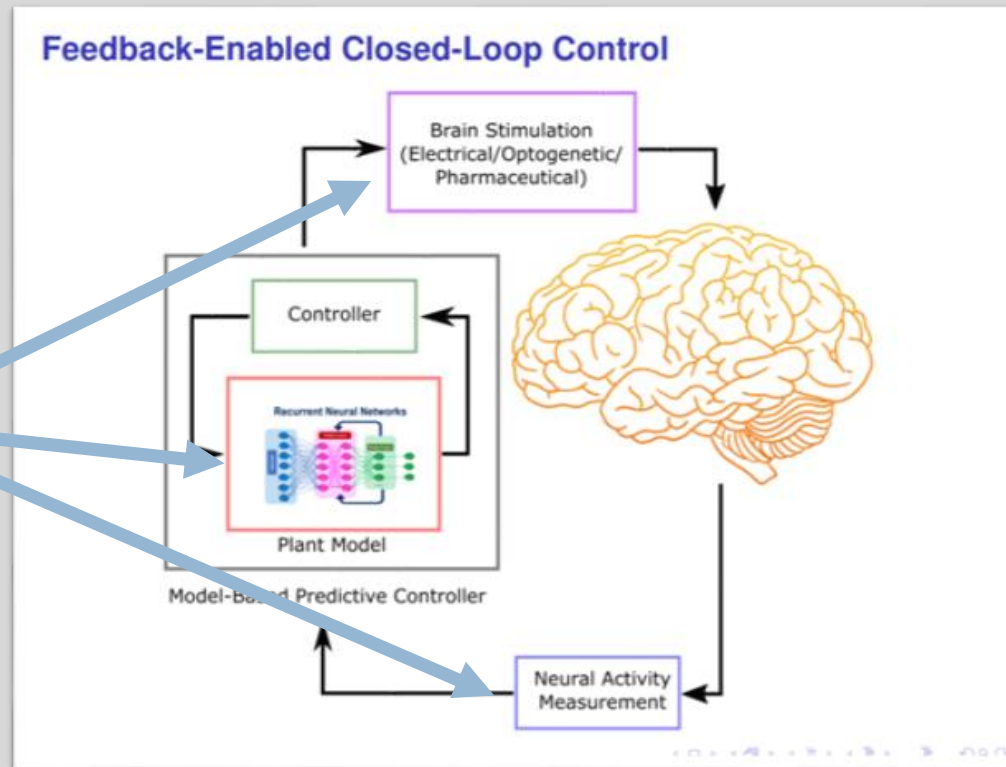
# Where are we?

Recomposition  
work partly  
contingent  
upon solutions  
in the controller  
segment

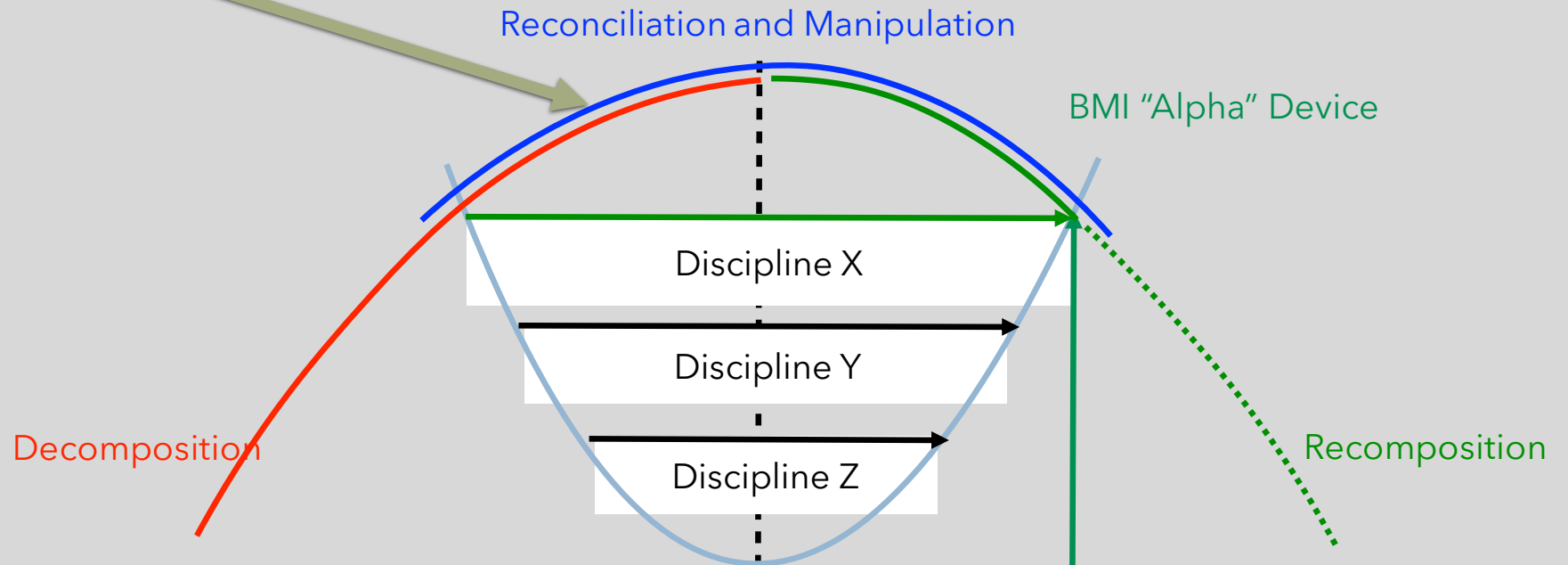


# Where are we?

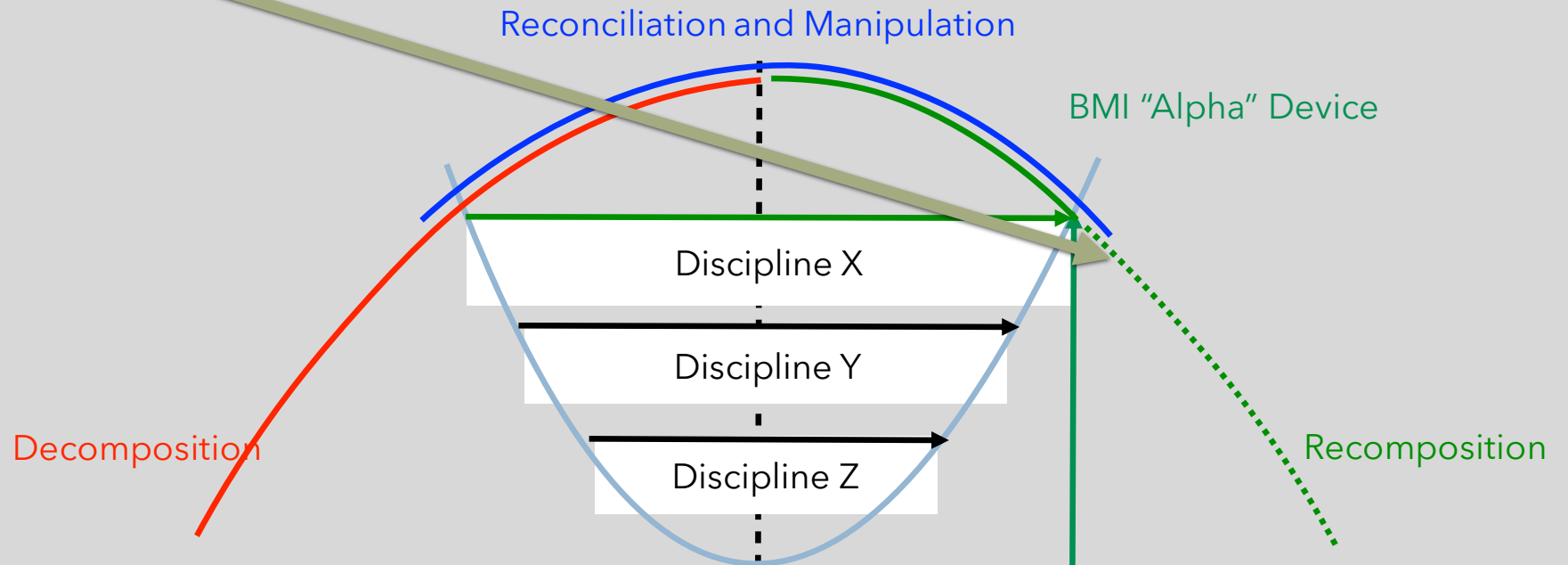
Many of the projects associated with the center have the character of reconciliation and manipulation work



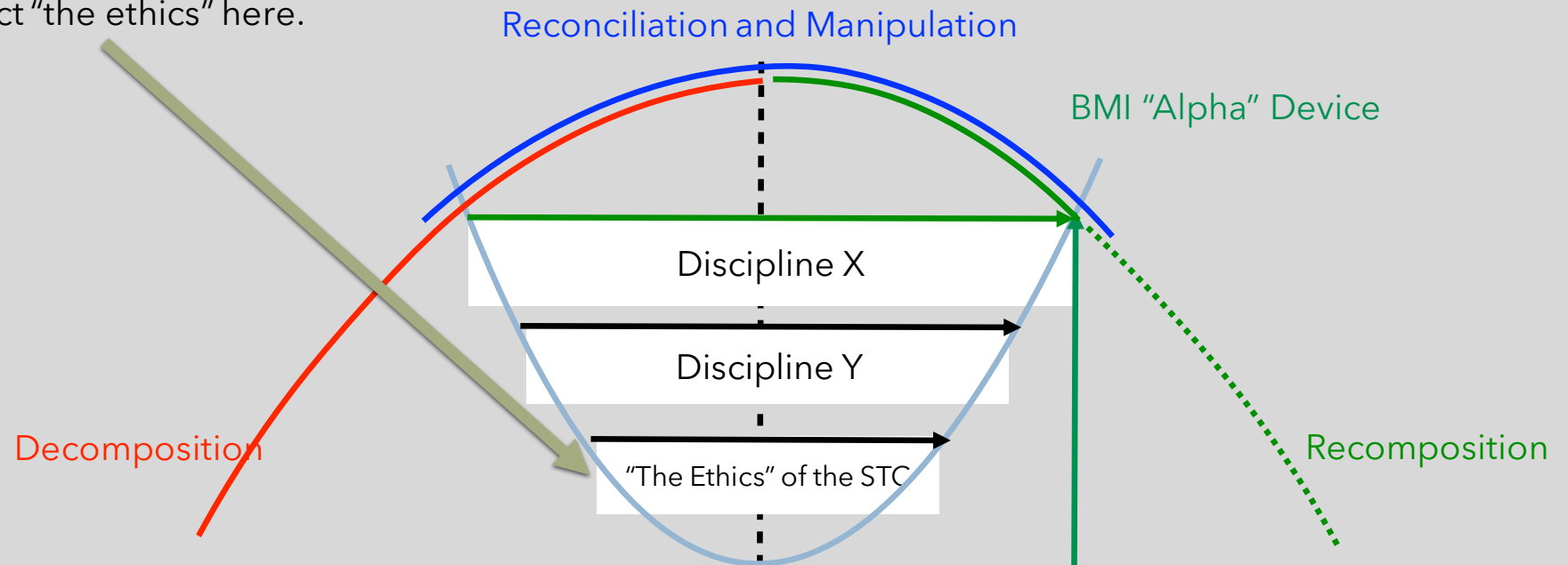
So, maybe we are here?



STC should move us here:



And the "workflow among workflows" means we construct "the ethics" here.



The background features a vibrant, abstract design. The top half is a solid orange color. Below this, a series of overlapping, semi-transparent geometric shapes—primarily triangles and circles—create a dynamic, layered effect. The colors of these shapes include red, purple, blue, green, and yellow, transitioning from left to right. The bottom half of the image is a solid light gray, providing a clean backdrop for the text.

## PART 2 SHORT-TERM G/B/R/W

# Analytical Core of "The Ethics"

- "The Ethics" of an engineering project needs to be about **outcomes** - the intended and unintended, chosen and avoided, *effects* of the relevant technology.
- Why? Engineering *isn't* (just) theory building.
- The contrast here is with (not outcomes but) pure theory or model building: the goal of "the ethics" for a STC should **not** be to, e.g., use BMI devices to try answer philosophical questions about the nature of ethical cognition.

# Analytical Core of "The Ethics"

- Adina Roskies (2001, 2021) distinguishes between:
  - Ethics of neuroscience
  - Neuroscience of ethics
- **Neuroscience of ethics** = using neuroscience to try to resolve classical problems in ethics - e.g. are all moral judgments really emotional projections? Perhaps limbic system activity correlated with occurrent moral judgements provides an answer.
- **Ethics of neuroscience** = meshing ethics and neuroscience together - e.g. "the ethics" of a BMI STC - e.g. a "workflow among workflows" - e.g. ethics as an organic part of the development of the neuroscience.
- Need an analytical core to "the ethics" that is squarely in the **ethics of neuroscience tradition** - its job isn't to help (even though they are important) neuroscience of ethics projects.
- That allows us to adopt (relative to standards in academic ethics!) a very simple conceptual framework as the analytical core of "the ethics".

# Analytical Core of "The Ethics"

I propose that we employ a simple 2x2x1 matrix of outcomes as the analytical core of "The Ethics" of the BMI STC.

	Probable Goods	Probable Bads	Neither
Probable Outcomes			
Improbable Outcomes			

Workflow = mapping outcomes, and then figuring out how to increase the probability of probable goods, reduce the probability of probable bads.

Doing **that** is "**the right thing to do**" for "the ethics" of a BMI STC, **relative to what is technically possible at the relevant stage of research**.

# Analytical Core of "The Ethics"

	Probable Goods	Probable Bads	Neither
Probable Outcomes			
Improbable Outcomes			

**More simply:** try to design-in probable(probably good) outcomes, and try to design-out any probable(probably bad) outcomes, so that they switch to improbable(probably bad) outcomes.

# Probably good? Probably bad?

- Treat “good” as a **cluster concept**:
  - **++Probability** (*x is good*) if there are some *w, y, z, ...* that are also probable goods and *w, x, y,* and *z* either tend to cause each other, or have a family resemblance with one another in terms of their causal powers, or both.
  - **--Probability** (*x is good*) if none of the conditions above are satisfied.
  - **Mutatis mutandis** for (*x is bad*)
- STC may support multiple “good outcomes” clusters, with high or zero overlap.
- Note that because good / bad treated as cluster concepts, pushing up or down the probability of some *x* should have network effects that, while not guaranteed to be uniformly good or bad, may allow for opportunistic ethical compounding.

# Probably good? Probably bad?

- Or, in plain English:
  - The probability that something is good goes up to the extent that it is in some kind of causal concert with other things that are probably good – good things just are probably good things that go together.
  - Ok to start with intuitions about whether or not something is good – the test is how well it meshes with other outcomes of interest – can use experiment to reject, modify, extend, blend, confirm, etc. hypotheses about probably good outcomes generated by, e.g., intuition (or creativity, or analogy, etc).
  - Threshold for determining something is “probably good” is same as the threshold for any other category being used in the project.

# Probable good? Probable bad?

- From a student's perspective, here's what "doing the ethical work" looks like.
  - List as many of the outcomes that intuitively have something to do with a patient's agency, autonomy, capacities, values, basic needs, practical fairness, privacy, cognitive uniqueness, physiological safety, and so on.
  - Next, for each outcome, assign it a probability that it can be caused **right now** given the stage of the technology.
  - For the outcomes that clear some threshold ( $Pr \geq 0.8?$ ), determine whether the outcomes are probable goods, probable bads, or neither **by investigating their potential for clustering**.
  - Note that asking "How could I cause clustering of these probable probable goods?" is in a very important sense **ethical engineering**, as it is figuring out how to deepen the network of *actually caused* good outcomes - these are the outcomes that matters!.
- Probable(probably good) outcomes should be **conserved** as research programme develops.

# Example #1

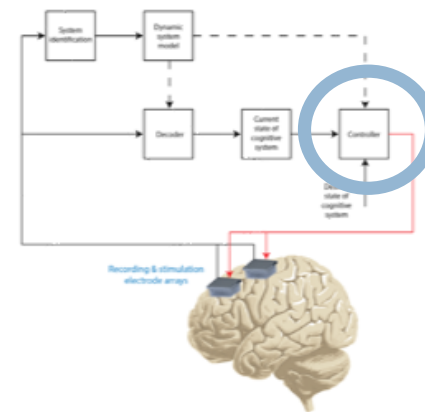
- From Dr. Zhaodan Kong (UC Davis)
- A probable(probably good) outcome that is a candidate for **conservation**.

# Where are we, redux?

Different options for inferring design of controller segment.


## System identification & control of neural cognitive systems

- Objective: To devise means for the precise modeling and control of neural activities pertaining to cognition
- Fundamentally, *how to control a system (i.e., the brain) with unknown dynamics*
  - CS: Learn Q function and/or control policy directly from data, i.e., RL
  - Control: Learn model first, then design a controller based on the model



# Probable Probably Good Outcomes

This is a good outcome!

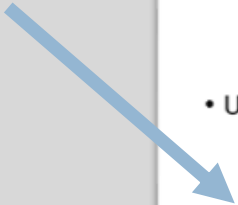


Main ideas: Control-theoretic modelling & model-based controller/decoder design

- Build a model that characterizes the effect of input stimulation on output neural activity/behaviors
  - Must quantify the time-evolution of neural activity/behavior
  - Amenable to controller/decoder design
  - Best choice: dynamical system models
- Utilize model-based controller/decoder design principles
  - Robust to noises and uncertainties
  - Principled, i.e., not based on heuristics and trials-and-errors
  - With theoretical guarantees therefore clinically safe

# Probable Probably Good Outcomes

This is a  
(research stage  
relative)  
(probably) good  
outcome!



Main ideas: Control-theoretic modelling & model-based controller/decoder design

- Build a model that characterizes the effect of input stimulation on output neural activity/behaviors
  - Must quantify the time-evolution of neural activity/behavior
  - Amenable to controller/decoder design
  - Best choice: dynamical system models
- Utilize model-based controller/decoder design principles
  - Robust to noises and uncertainties
  - Principled, i.e., not based on heuristics and trials-and-errors
  - With theoretical guarantees therefore clinically safe

# Probable Probably Good Outcomes

Does using  
(e.g.) recurrent  
neural networks  
in controller  
segment cost as  
the ability to  
bake-in certain  
theoretical  
guarantees?

## A wide range of $f$ s

Increasing complexity

- |  |  |  |
|--|--|--|
| <ul style="list-style-type: none"><li>• Linear dynamical system models<br/><math>f(x, u) = Ax + Bu</math></li><li>• Pros:<ul style="list-style-type: none"><li>• Easy to control and learn</li><li>• A wide range of applicable analysis and controller design tools</li></ul></li><li>• Cons:<ul style="list-style-type: none"><li>• Relevant dynamics possibly nonlinear</li></ul></li></ul> | <ul style="list-style-type: none"><li>• Recurrent switching models<br/><math>f(x, u) = \{A_i x + B_i u, x \in X_i\}_{i=1}^M</math></li><li>• Pros:<ul style="list-style-type: none"><li>• Can, in principle, approximate any nonlinear system</li><li>• A wide range of applicable analysis and controller design tools</li></ul></li><li>• Cons:<ul style="list-style-type: none"><li>• May still need a large <math>M</math> if the underlying dynamics is quite nonlinear</li></ul></li></ul> | <ul style="list-style-type: none"><li>• Deep neural networks, such as LSTM and RNN</li><li>• Pros:<ul style="list-style-type: none"><li>• Capture highly complex and nonlinear activities</li></ul></li><li>• Cons:<ul style="list-style-type: none"><li>• Need lots of data and/or expertise to design the NN structure</li><li>• Extremely hard to control and analyze (many open theoretical and practical questions)</li></ul></li></ul> |
|--|--|--|

# Probable Probably Good Outcomes

Yes!

So, adopt as an ethical rule: no “opaque” ML or neural nets in the controller segment.

## A wide range of $f$ s

Increasing complexity →

- |  |  |  |
|--|--|--|
| <ul style="list-style-type: none"><li>• Linear dynamical system models<br/><math>f(x, u) = Ax + Bu</math></li><li>• Pros:<ul style="list-style-type: none"><li>• Easy to control and learn</li><li>• A wide range of applicable analysis and controller design tools</li></ul></li><li>• Cons:<ul style="list-style-type: none"><li>• Relevant dynamics possibly nonlinear</li></ul></li></ul> | <ul style="list-style-type: none"><li>• Recurrent switching models<br/><math>f(x, u) = \{A_i x + B_i u, x \in X_i\}_{i=1}^M</math></li><li>• Pros:<ul style="list-style-type: none"><li>• Can, in principle, approximate any nonlinear system</li><li>• A wide range of applicable analysis and controller design tools</li></ul></li><li>• Cons:<ul style="list-style-type: none"><li>• May still need a large <math>M</math> if the underlying dynamics is quite nonlinear</li></ul></li></ul> | <ul style="list-style-type: none"><li>• Deep neural networks, such as LSTM and RNN</li><li>• Pros:<ul style="list-style-type: none"><li>• Capture highly complex and nonlinear activities</li></ul></li><li>• Cons:<ul style="list-style-type: none"><li>• Need lots of data and/or expertise to design the NN structure</li><li>• Extremely hard to control and analyze (many open theoretical and practical questions)</li></ul></li></ul> |
|--|--|--|

# Further Examples of Probable (Probably Good) Outcomes [speculative]

- **Controller algorithm neutrality with respect to functional architecture of cognition** – if the goal ultimate goal is “rational decision detection, transmission, and re-transmission” then likely don’t need to build into the controller something that is functionally isomorphic to cognition. (Probably) supports autonomy and user diversity, maybe also existential privacy (no ‘mirror of the soul’ in the telemetry).
- **User-in-the-loop at controller segment** – calibration is a fact of life, and “a decision” (the signal?) is something that can be user-verified and thus a user-in-the-loop process could be a necessary tool for BMI-device-to-unique-brain calibration. *Where feasible*, (probably) supports aligning use cases with user/patient values, and autonomy; with attention to recruiting may also support diversity.
- **Robustness to Cognitive Differences** – much in the way you want your car or plane to be operational in the widest range of environment scenarios, important to develop technology with an eye to ensuring it is functional across the widest possible set of use cases. (Probably) supports trustworthiness (How can we tell *ex ante* that person A is relevantly different than person B?, diversity, and affordability).

Key Intuition: These can be design objectives in the short-term. *We could try to make these probably good outcomes.*

# Examples of Probable (Probably Bad) Outcomes [speculative]

- **Transformative Experience Risk** – we are starting to understand that BMI devices can cause “transformative experiences” – some users of certain devices seems to be a categorically different person before and after implantation. Investigating techniques and processing for disclosing this in advance is likely obligatory – but ameliorating seems also to be goal. “Transformative experiences” should not be side-effects.
- **Steep Performance Drop-off Curve** – for implantables, need a process for ensuring durable function, maintenance, and potentially removable – process too for defining a floor of performance. Users should bear the risk of be platforms for short-term scientific breakthroughs that are abandoned after spotlight moves on.

Key Intuition: These too could be design and process objectives in the short-term. *We could design to avoid (probably) bad outcomes.*

# Short-Term STC Ethics

Short Term Goal = Try to avoid causing probably bad outcomes -- turn probable(probably bad) into improbable(probably bad).

	Probable Goods	Probable Bads	Neither
Probable Outcomes	++++++	----	....
Improbable Outcomes	----	++++	

The **right** thing to do, then, is to maximize to the extent possible the *probability* of the probable goods (considered as a cluster or network), while also minimizing to the extent possible the *probability* of the probable bads (considered as a cluster or network).

# Short-Term STC Ethics

	Probable Goods	Probable Bads	Neither
Probable Outcomes	++++++	----	....
Improbable Outcomes	----	++++	

So we have an “ethics”: a framework that defines the good, bad, right, and wrong, and tells us that we learn about what falls into these categories by paying close attention to the properties of the BMI technology as it develops.

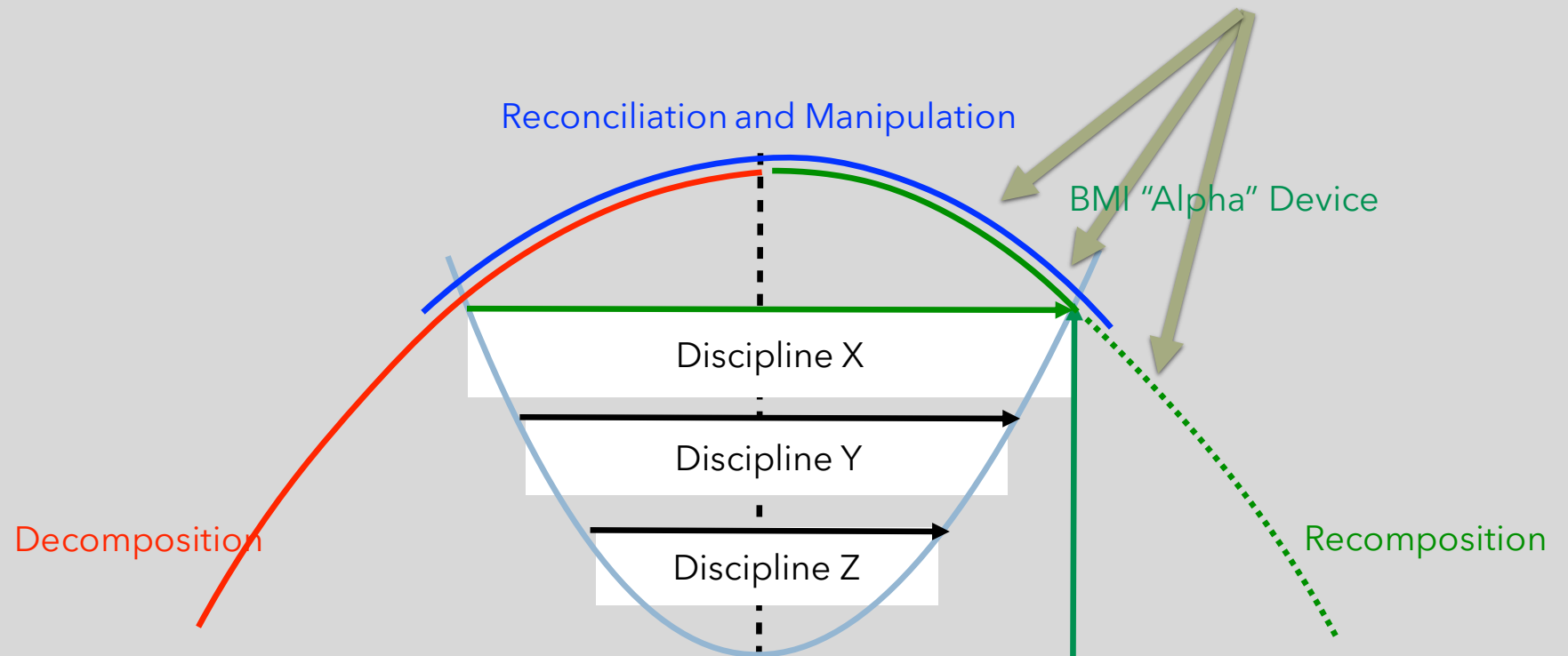
The background features a vibrant, abstract design. The top half is a solid orange color. Below this, a series of overlapping, semi-transparent geometric shapes—primarily triangles and circles—create a dynamic, layered effect. The colors of these shapes transition from warm tones (red, orange, yellow) on the left to cooler tones (blue, green, white) on the right. The bottom half of the image is a solid white background.

# PART 3: MEDIUM-TERM G/B/R/W

# Short-Term to Medium-Term

- Transition from Short-Term to Medium-Term likely is anchored on when “BMI Device Alpha” can be designed.
- Note that this is going to produce a **totally new** set of outcomes, many of which we might not be able to forecast now.
- Likely implies we would have to retire the “Short-Term” Table of Good and Bad Outcomes.
- **But it matters at this point that we know what outcomes we are trying to conserve - the ability to produce certain probably good outcomes and mitigate/avoid certain probably bad outcomes.**

Transition Point from Short-Term to Medium Term?



# Same Heuristic Applies after Transition Point

Try to avoid causing probably bad outcomes.

	Probable Goods	Probable Bads	Neither
Probable Outcomes	++++++	----	
Improbable Outcomes	----	++++	

**As before:** The right thing to do, then, is to maximize to the extent possible the *probability* of the probable goods (considered as a cluster or network), while also minimizing to the extent possible the *probability* of the probable bads (considered as a cluster or network).

# Same Heuristic + Public Goods & Bads

- But we need to change the definition of “probable good/bad” to account for broader use of the technology.
- In ethics and political philosophy, the relevant concept that gets added here is that of a **public good**.
- Similar to what economists call **welfare** (but not what they call public goods!).
- Public goods to care about: inclusiveness, equitability of access, acceptable price discrimination (maxi-min?), take end-user capabilities as input.
- At least some of these good outcomes will not be the effects of device-level properties – they will be regulatory, economic, political, or legal.
- But they are still (possible) clusters – it just is (much!) harder to create the conditions conducive to clustering.
- So, this is a more complicated **analytical project** than organizing (e.g.) short-term probable(probably good) outcomes – but it is a simpler task if we prepare the ground for this work.

# Transition: Conserving Legacy G/Bs

- **Implication:** possible to adopt a meta-principle of **conservation**: to the extent possible, preserve the ability of the technology to cause the “probable goods” of the reconciliation and manipulation stage.


Short Term	Probable Goods	Probable Bads
Probable Outcomes	++++++	----
Improbable Outcomes	----	++++

Medium Term	Probable Goods	Probable Bads
Probable Outcomes	+++++++ +	--
Improbable Outcomes	--	+++++++

# Transition: Conserving Legacy G/Bs

- **Implication:** possible to adopt a meta-principle of **ethical conservation**: to the extent possible, preserve the ability of the technology to cause the “probable goods” of the reconciliation and manipulation stage.

Short Term	Probable Goods	Probable Bads
Probable Outcomes	+++++	----
Improbable Outcomes	----	++++




Medium Term	Probable Goods	Probable Bads
Probable Outcomes		
Improbable Outcomes		

# Transition: Conserving Legacy G/Bs

- **Implication:** possible to adopt a meta-principle of **ethical conservation**: to the extent possible, preserve the ability of the technology to cause the “probable goods” of the reconciliation and manipulation stage.

Short Term	Probable Goods	Probable Bads
Probable Outcomes	+++++	----
Improbable Outcomes	----	++++



Medium Term	Probable Goods	Probable Bads
Probable Outcomes	++	
Improbable Outcomes		+

# Transition: Conserving Legacy G/Bs

- **Implication:** possible to adopt a meta-principle of **ethical conservation**: to the extent possible, preserve the ability of the technology to cause the “probable goods” of the reconciliation and manipulation stage.

New Short Term	Probable Goods	Probable Bads
Probable Outcomes	++	
Improbable Outcomes		+

# End-Point of Medium-Term Ethics

	Probable Goods	Probable Bads	Neither
Probable Outcomes	Public Goods and "Legacy Goods"	----	...
Improbable Outcomes	----	Public Risks, Sociotechnical Risks and "Legacy Risks"	

**Still:** The right thing to do, then, is to maximize to the extent possible the *probability* of the probable goods (considered as a cluster or network), while also minimizing to the extent possible the *probability* of the probable bads (considered as a cluster or network).

The background features a vibrant, abstract design. The top half is a solid orange color. Below this, a series of overlapping, semi-transparent geometric shapes—primarily triangles and circles—create a dynamic, layered effect. The colors transition from orange and red at the top, through purple and blue in the middle, to green and light yellow at the bottom. The shapes are arranged in a way that suggests a mountain range or a series of peaks. The text "PART 3: PRACTICALITIES" is centered in the lower half of the image, overlaid on the white background.

## PART 3: PRACTICALITIES

# "The Ethics" as a Workflow

- Remember where we started - the idea that "the ethics" of the STC should be a workflow among workflows.
- How do we implement the proposed conceptual framework?
- Two practical questions: **training** and a process for **doing "the ethics"**.

# "The Ethics" -- Training

- **Training:**

- **Preliminary:**

- How cluster concepts work
    - Case studies of ethics-by-way-of-cluster-concepts in (e.g.) developmental economics and healthcare process formation.
    - Concept formation for concepts like agency, autonomy, capacity, choice, transformative experience, individual risk, privacy, cognitive diversity, etc. – by way of case studies and literature review.

- **Ongoing:**

- Cross-training in neuroengineering (what is the state of the tech?) and allied fields like anthropology, sociology, political economy, and so on.

- **At Transition from Short-Term to Medium-Term**

- Additional round of concept formation focusing on moral concepts not linked to individuals but, instead, public groups: e.g. price discrimination, bias, liability, inclusiveness, fair use, etc.

# "The Ethics": Process during Stage 2

- Process:
  - **Annual "Ethics Afternoon"**
    - Structured as a regular academic retreat with talks, group exercises, Q&A, etc.
    - Provide **Preliminary Training**.
    - For new members of the STC, focused on entering graduate students.
    - Fall in Davis is lovely!
  - **Outcomes Analysis Group**
    - Meets once a month, attendance semi-mandatory, with set agendas, capacity to form subcommittees, and so on.
    - Structured on the model of an academic medical center's clinical ethics case review committee
      - Cross-disciplinary team of specialists who used a simple shared conceptual framework to evaluate emerging technologies, complicated use cases, and who provide expertise-led cross-training where appropriate.
    - Provides **Ongoing Training** and **BMI Alpha Device Transition Forecasting**
    - Requires ongoing involvement from STC PIs – ensuring that "ethics" and "science" and "engineering" remain in sync.
    - Goal is to support the long-term goals of the STC by "populating" the short-term and medium-term tables of probable-probable goods/bads.
    - Provides a capacity to deal with emergent issues – e.g. unforeseen risks.

# "The Ethics" Process at Transition from Stage 2 to Stage 3

- **Process:** When pre- to post-BMI Alpha Device seems likely, form a new subgroups of **Outcome Analysis Group** and add an additional annual Ethics retreat
  - **Legacy Outcomes, Public Goods, and Public Risks Working Group**
    - **Not** structured on the model of a clinical ethics workgroup.
    - **More engineering, policy, and law than pure ethics** – tasked with solving the design problems of (to the extent possible) **baking in** the probable probable goods and improbable probable bads in the post-BMI Alpha device technology.
  - **Annual Public Goods, Public Risks and BMI Devices Retreat**
    - Structured as a regular academic retreat with talks, group exercises, Q&A, etc.
    - Provide **Concept Formation** in clusters centered on public goods and wide-spread capabilities, regulatory realities, etc.
    - For all members of the STC, but focusing on brining in “public goods” expertise – e.g. lawyers, regulators, economics, user-advocacy groups .
    - *Late Fall in Davis is lovely!*

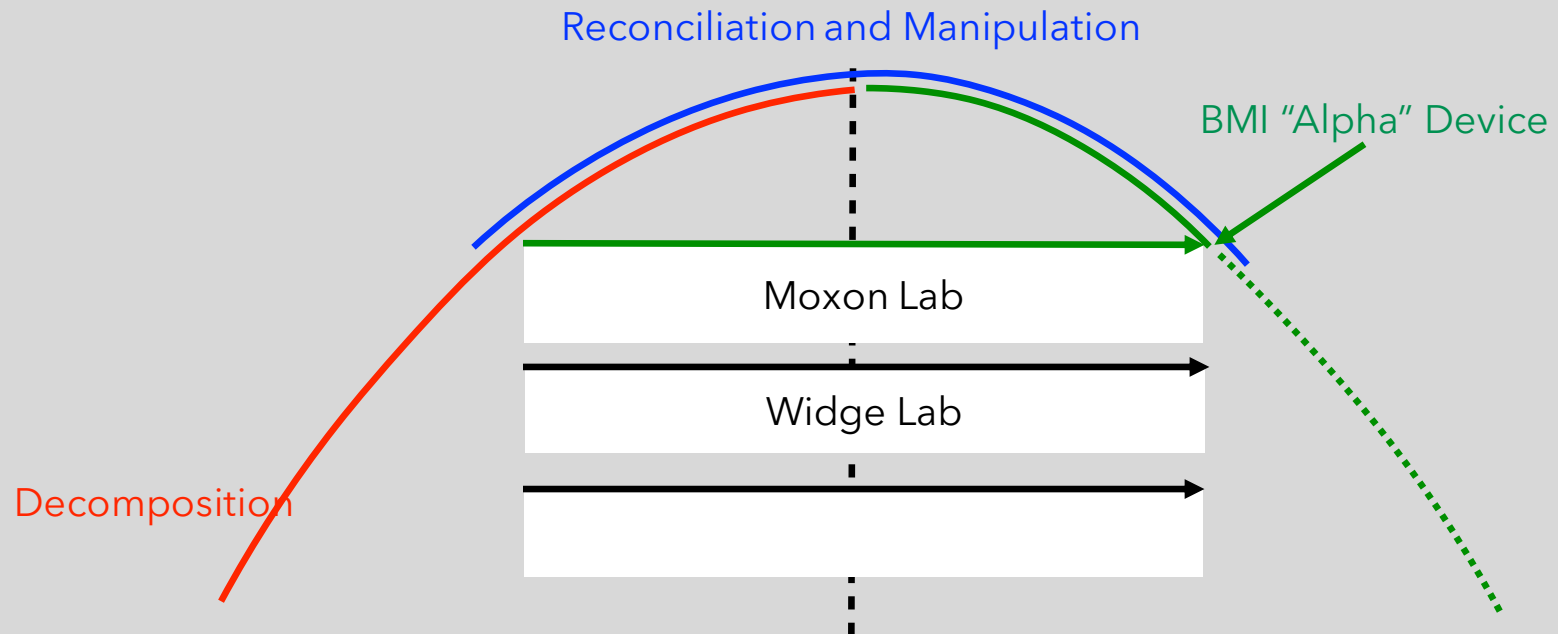
The background features a vibrant, abstract design. The top half is a solid orange color. Below this, a series of overlapping, semi-transparent geometric shapes—primarily triangles and circles—create a dynamic, layered effect. These shapes are colored in a spectrum including red, purple, blue, green, and yellow. The bottom half of the image is a plain white background. The text "PART 4: CONCLUSION" is centered in the white area.

## PART 4: CONCLUSION

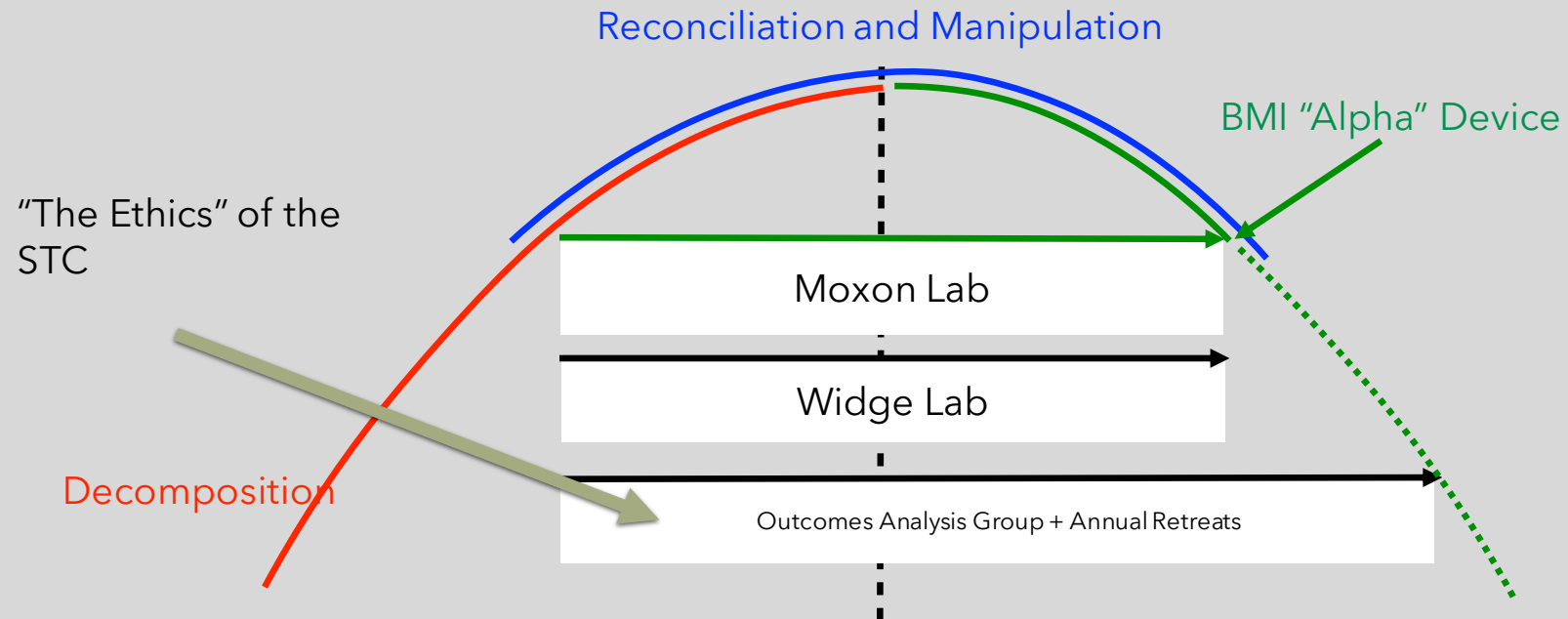
# "The Ethics" of a BMI STC

- Identify probable (probably good) outcomes and probable (probably bad) outcomes, and using the result of this work to shape the design and use parameters of BMI Device Alpha.
- Configured as a "workflow among workflows".

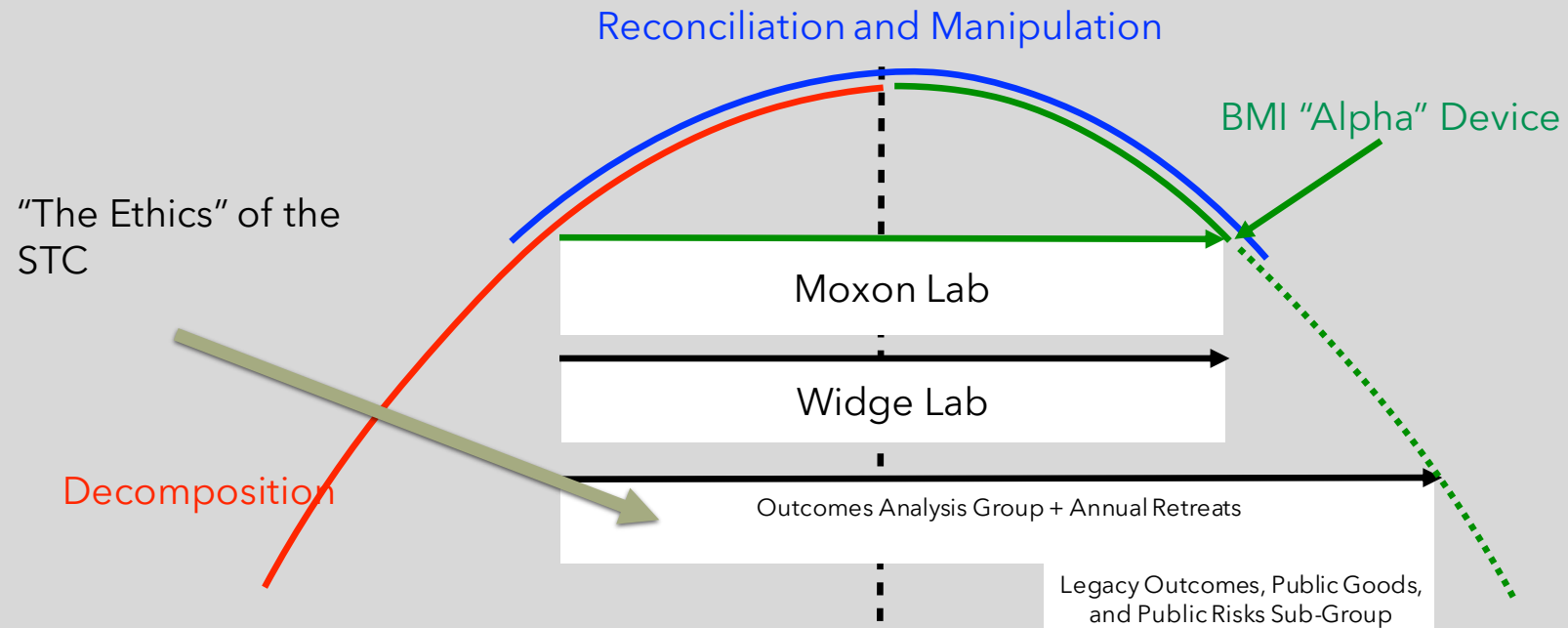
# "The Ethics" of a BMI STC as a Workflow among Workflows



# "The Ethics" of a STC as a Workflow among Workflows



# "The Ethics" of a STC as a Workflow among Workflows



# The Work of "The Ethics" of a STC

Reduced to its simplest: first step is to build a team whose job is to populate this table, in concert with the technology as its development progresses, and as the STC itself develops, explores ways to conserve the good outcomes.

	Probable Goods	Probable Bads	Neither
Probable Outcomes			
Improbable Outcomes			