

# Punk as ideology

Gavin

2010-09-06

```
{% include punk/links.md %}


<!-- Church of Euthanasia 'activist', c.2005. --&gt;
&lt;br&gt;&lt;br&gt;
&lt;blockquote&gt;
    Punk is a music that is inherently hilarious. To try to make serious punk music is like
&lt;/blockquote&gt;
&lt;p style="line-height:5px"&gt;-&lt;a href="{{reflux}}&gt; my mate James&lt;/a&gt;
&lt;/p&gt;
&lt!-- --&gt;
&lt;br&gt;
&lt;blockquote&gt;
    It has always been my way to de-value the fashionable, light-hearted, impulsive traits t
&lt;/blockquote&gt;
&lt;p style="line-height:5px"&gt;- Greg Graffin&lt;/p&gt;</pre>
```

Punk is ancient: say 40 years old. Anything this old in the modern age will in fact be several different things uneasily sharing a name. 3

It's easy to view punk as necessarily political and necessarily left-wing. But neither are generally true. Just one instance before we get into the weeds: The Ramones, the *central* example of a punk band, have exactly one even vaguely political song, a mild response to one faux pas by Reagan. (Their guitarist was a crabby Republican.)

At first, punk was negative: defined by its opposition and not a positive programme; and also as in nihilistic and downbeat, focussing on the worst things in the world for aesthetic reasons.

For best effects, set this playlist going before you continue.

```
{% include punk/history.md %}
```

So it wasn't political, and then it was.

## The ideology

Empty factories to the east and all our waste  
The shape of things that came,  
shown on the broken worker's face  
To the west you'll find a silicon promised  
land Where machines all replace their minds with systematic profit plans  
The course of human progress staggers like a drunk Its pace is quick and heavy but  
its mind is slow and blunt I look for optimism but I just don't know Its seeds  
are planted in a poison place where nothing grows Just one political song, just  
one political song To drop into the list that stretches years and years long Just  
one political song, just one political song To drop into the list that stretches  
years and years long

– Op Ivy (1989)

One way to handle the messiness and variance of political ideologies is to give up the search for sufficient beliefs, away from the attempt to cleanly distinguish true punks from fake punks. Instead you can try splitting it into

- core beliefs (the essence of the thing),
- adjacent beliefs (other commonly found themes),
- peripheral (fringe but distinctive) beliefs.

This chunky approach lets us account for the mind-jarring variety of people that all call themselves liberals, or socialists, or whatever. The downside is that this always makes the “core” sound banal, because it has to be something that appeals to everyone.

```
<h3>A first attempt</h3>
<div>
  <ul>
    <li>Authenticity</li>
    <li>Social conscience</li>
    <li>Alienation</li>
    <li>Freedom</li>
  </ul>
  <!-- -->
  There are systematic opposite trends for each of these: e.g. Sid's insincere swastika sh
</div>
<h3>Situation</h3>
<div>
  A standard theory of punk equates it with <a href="{{sit}}>Situationism</a>, French abs
</div>
```

## Core elements

- *Anti-elitism*. Entails amateurism, hatred of pretence. (This is *not* the same as egalitarianism, below.)
- *Alienation*. From something or everything mainstream: careerism, con-

sumerism, ordinary social interactions, romance, nation, foreign policy, religion, species, even hedonism. In particular:

- *Anti-establishment*. The System. If you have the energy you might dream of breaking away and creating a perfect part inside the evil whole, the punk house.
- *Individualism*. Entails nonconformity; customized, deviant appearance [itself a new uniform]; drugs; refusal of ordinary roles. Free thought. 4

### Adjacent elements

- *Snobbery*. A rejection of pop culture, as brainwashing commercialised shit or just shit. One in a series of rejections one must conduct to be truly DIY. (Entails abhorrence of popularity, the fear and hatred of “selling out”. Often painted as anti-consumerism.)
- *Inverse snobbery*. Off-hand rejection of bourgeois quality and tradition. Generally not because of a Marxist critique, but because boring or inaccessible or old. (This is just anti-elitism again.) 5
- *Authenticity*. Honesty and autonomy as far more important than quality. DIY and amateurism and abhorrence of profit.
- *Irony*. A lot of punk is self-consciously dumb fun, and a lot of the rest covers extreme things for the hell of it.
- *Nihilism / Pessimism*. Focus on the worst things in the world. War, crime, atrocities, etc. Often identification with the freakish, the low, the cheap and the taboo, for kicks (Lou Reed, Iggy) or semi-political art (Patti Smith). Obsession with kitsch and cool. (Richard Hell) Sometimes also primitivism.
- *Egalitarianism*. But often in a specific inverted way: “We are all sub-human scum. You’re no better than me, I no more than you.”. Entails the DIY ethic, bands playing in the crowd - but more importantly an usual degree of activism, direct action, civil disobedience.

look at you, then look at me / there is no difference I can see

– D.R.I, rather than say Barney the Dinosaur

- *Internationalism*. One reason punk remains so lively is that people listen to foreign bands, and then somehow manage to book them for international tours. This makes the community much larger, more interesting, and more sustainable. Among the few other cultures which pull this off to this degree are metal, Art and football.
- *Anti-capitalism*. Common but not as common as you’d think from outside. Few bands really took it on with real seriousness. Entails guilt

about consumerism, disdain for material comforts, opportunistic squatting, freeganism. 6

- *Rationalism.* Atheism, Skepticism, lip service to the great social theorists. This is the upside of basing your worldview on Noam Chomsky's.
- *Animal rights & environmentalism.* I don't know numbers, but definitely a disproportionate awareness of social justice, ethical and environmental issues, at least where this conforms to the above anti-state, anti-tradition rules. (Entails veganism and BIKEPUNK.)
- *Syncretism:* basically any genre has some band doing the “-punk” version of it. The ska-reggae-dub-punk complex of the early 80s was particularly important. This is the antidote to the purists still listening to dull iterations of exactly the same music forty years on. I think this is yet another factor in punk's longevity. 7

### Peripheral elements

MiSanDao, Chinese oi

- Working-class consciousness: As in far-left Oi.
- *Racism:* As in far-right Oi. Apart from the explicit kind, the rest is maybe just the background level of society at large, or maybe exacerbated by the edginess and dark irony. Punk, like its descendent indie music, is disproportionately white, but note that this may be a boring statistical thing.
- *Feminism:* There was a bit of gender flex in punk originally (compare Joey Ramone to Mick Jagger), and many women in the NY mix. But, then hardcore was *much* more macho than ordinary society. It took a while for the natural mix of feminism and punk to really show up.
- *New Anarchism.* Anarchism was fairly dormant from the end of the Spanish Civil War to the 70s. While it was previously a thing for angry young men, it is now the quintessential youth politics, away from the mannered, literary resistance of Kropotkin or Proudhon, towards the tragically hip Hakim Bey and Howard Zinn and David Graeber.
- *Libertarianism.* I have in mind people like Steve Albini and Frank Kozik), but you see this in the anti-SJW kind of punk a lot - “you can't judge me”.
- *Straight edge.* Self-control fetishism. No booze, no drugs, often no meat, and sometimes even no casual sex.

‘Punk rock’ is a word used by dilettantes and heartless manipulators about music that takes up the energies, the bodies, the hearts, the souls, the time and the minds of young men who give everything they have to it.

- Iggy Pop

It is a style, it became an ideology, and it remains a community. All in all, a good way to spend your teens.

{% include punk/against.md %} {% include punk/leatherface.md %} {% include punk/epigram.md %}

{% include punk/foots.md %}

# Showing over saying

Gavin

2010-09-10

```
{% assign t = "https://en.wikisource.org/wiki/Tractatus_Logico-Philosophicus"
%} {% assign white = "https://www.goodreads.com/book/show/776361.Wittgenstein_s_Tractatus_Logico_Phi
%} {% assign irony = "https://absoluteirony.wordpress.com/2014/09/17/nagarjuna-
nietzsche-rorty-and-their-strange-looping-trick/" %} {% assign body =
"https://people.umass.edu/klement/tlp/tlp.html#bodytext" %}

{% include tractatus/update.html %}
{% include tractatus/preamble.html %}
{% include tractatus/setup.html %}
```

---

## 1. The world is all that is the case.

1.13

## 2. What is the case is the existence of states of affairs

```
<h3>TODO</h3>
<div>
    2.12 - A picture is a model of reality.<br />
    2.141 - A picture is a fact.<br />
    2.172 - A picture cannot depict its pictorial form: it displays it.<br />
    2.19 - Logical pictures can depict the world.<br />2.223 - In order to tell whether a p
    2.224 - It is impossible to tell from the picture alone whether it is true or false.<br />
</div>
```

## 3. A logical picture of facts is a thought.

```
<h3>TODO</h3>
<div>
    3.01 - The totality of true thoughts is a picture of the world.<br />
```

```
    3.1 - In a proposition a thought finds an expression that can be perceived by the senses  
    3.3 - Only propositions have sense; only in the nexus of a proposition does a name have  
    3.332 - No proposition can make a statement about itself, because a propositional sign  
  </div>
```

#### 4. A thought is a proposition with a sense.

4.001

```
<h3>TODO</h3>  
<div>  
    4.003 - Most of the propositions and questions to be found in philosophical works are no  
    4.0031 - All philosophy is a 'critique of language'. The apparent logical form of a prop  
    4.11 - The totality of true propositions is the whole of natural science<br />  
    4.461 - Propositions show what they say; tautologies and contradictions show that they s  
    4.464 - A tautology's truth is certain, a proposition's possible, a contradiction's impo  
</div>
```

#### 5. A proposition is a truth-function of elementary propositions.

```
<h3>TODO</h3>  
<div>  
    5.3 - All propositions are results of truth-operations on elementary propositions.<br />  
    5.6 - The limits of my language mean the limits of my world.<br />  
    5.61 - We cannot think what we cannot think; so what we cannot think we cannot say either  
    5.621 - The world and life are one.<br />  
    5.63 - I am my world. (The microcosm.)<br />  
    5.632 - The subject does not belong to the world but it is a limit of the world <br />  
</div>
```

#### 6. Truth-functions and propositions have the same general form

6.02 6.54

```
<h3>TODO</h3>  
<div>  
    6.13 - Logic is not a body of doctrine, but a mirror-image of the world. Logic is trans  
    6.2 - Mathematics is a logical method.<br />  
    6.21 - A proposition of mathematics does not express a thought.<br />  
    6.41 - The sense of the world must lie outside the world.<br />  
    6.431 - At death the world does not alter, but comes to an end.<br />  
    6.4311 - Death is not an event in life: we do not live to experience death.<br />6.44 -  
</div>
```

#### 7. Whereof we cannot speak, thereof we must be silent.

All the things in these pictures can easily be said: they all fit into a set of propositions (albeit a self-refuting set): it's called Tractatus Logico-Philosophicus.

Drawing is only a kind of saying; the kind of reference that he means by zeigen is less concrete and more important than this.

# ‘Have One on Me’ (2010) by Joanna Newsom

Gavin

2010-10-30

```
{% assign hook = "http://www.jstor.org/sici?sici=0261-1430%28198701%296%3A1%3C1%3AATO"
%} {% assign bon = "https://en.wikipedia.org/wiki/Bonnie_%22Prince%22_Billy"
%}
```

... do her words really need to be broken down like formulae? I think not. Simply to escape into the world of Joanna and be encapsulated into it and applauding it is enough. And maybe not understanding completely is the more beautiful act of musical appreciation, lack of total understanding leaves the listener with a humbled nice sense of ignorant awe.

— Guardian commenter

(Forgive me; I am not satisfied by ignorant awe.)

## Give love a little shove and it becomes terror

Newsom is hard work. Well, Have One On Me is a map of the heart, and you shouldn't expect those to offer themselves lightly.

The album covers the forms of love: divine or agape (tracks (3, 7, 14); filial (track 9, 14); courtly (track 2); obsessional (tracks 1, 5, 10); maternal (track 6! but touches in 1, 5, 11, 14); platonic (passionate friendship: track 8 and maybe 11); panicked (track 4); dependent (track 5, 10, 16); wilful (track 1, 16); of place (track 9); destructive (2, 8, 10, 15, 16?, 17?); forbidden (track 2); unrequited (track 18 above all, but 1, 7, 10, 15) and love of self (track 3, 4, 13). It exalts, despairs, casts about in the land.

Rock reviews miss the point in territory like this. There was a great deal written about it being a triple!! album!!!, which obscures the real way it's ambitious; this 123-minute thing requires patience because of its richness, not its length. The length (songs 6 mins on average), her vocabulary, voice, caesura, unfamiliar instruments slow us down, and then there's the alien allusions that leave us out, first of all.

Pretension, affectation, whimsy are just unavoidable side-effects of ambition. The lyrics work on their own as poetry, which is so rare in even the best pop music 1.

It ain't Renaissance music, but it is sacred. (American Secular Sacred). My mate James says it's "*a book of an album. It's Middlemarch*", and this is the case. Though, since it's episodic and woozy and dark, I'd call it Nabokov's Ada more. James also spits at people who emphasise the bits of her that appear Medieval - but the fact is, she is making historical music; it's drenched in dead music. But it's the blues; Ol' Opry cakewalks; cabaret; parlour-music; Appalachiana; and gospel, rather than the pre-Baroque. (Gershwin > Gibbons.) Given this marinade of early American popular music and William Faulkner, Newsom sounds lasting.

I don't listen to her for historical satisfaction. What I love about it are the many moments of perfect sound and sense, the grand hooks". There's so many here because the songs are so long and get the time to climb all that way up.

## Man vs Life

A type of love pointedly missing in the above rundown is empathic love. Where Ys burbled with anthropomorphisations, companion animals, and a general affinity with the universe, Have One on Me, while still full of nature, is much more about the Rancher (a lonesome, domineering social product nestling in a hostile world). It's sensual, snug, and macabre where Ys was abstract, epic, and pure.

"I hope Mother Nature has not overheard!(Though, she doles out hurt like a puking bird.)" - You and Me, Bess

"Driven through with her own sword, Summer died last night, alone." - Autumn  
"Wolf-spider, crouch in your funnel nest,... have I had a hand in your loneliness?"  
- Go Long

"Black nose of the dog / As cold as a rifle " - Ribbon Bows

With nature so terrible, the only safe place is civilisation, specifically the arms of someone who may or may not stay. The cover is filled with dead things: a judgmental peacock, half-plucked; a stuffed deer wearing a feather headdress; a divan draped in leopardskin - and her, langorous and deathly in the centre. And her animal motif-characters are this time uniformly malign - even Bess the horse makes "glad neighing", at highwayman-Joanna's hanging.

The significance is that the animals are aspects of the human characters. Newsom deals with the coldnesses, stubbornesses or malices of the male lead and female lead via animal symbols.

```
<h3>Motifs</h3>
<div>
  <ul>
    <li>"WATERS" ( which both separates and connects two banks, or, fertility)</li>
    <li>THE DANGERS OF FEMININITY
      11. "...my ankles are bound in gauze, sickly dressage,"<br />
```

```

16. "My mama may be ashamed of me with all of my finery..."<br />
18. "I have gotten into some terrible trouble / beneath your blank and rinsing g
</li>
<li>GOD IS SHIT, the indifference of Nature<br />
    2. "like a cornered rat"<br />
    10. "my faith makes me a dope"<br />
    13. "I glare and nod, like the character, God bearing down"<br />
    14. a feared mistress<br />
    16. using your dog as your theologian<br />
    16. "When I am alone, I take my god to task"<br />
    17. "To whose authority do you consign your soul?"<br />
</li>
<li>
    FLAME / BURNING (that feel.)<br />
</li>
<!-- + TRAVEL & EXILE; EAST vs WEST<br /> -->
<!-- + "BLUE" -->
</ul>
</div>

```

Henri Rousseau, “The Dream” (1910)

## The arc

The best hope for a unified story arc comes if we pick out the farm couple, seen most clearly in track 5, “No Provenance”. This easily ties into the Californian childhood arc, which is also the one who is intrigued by Lola Montez and empathises with her vengeance. My reading splits things into:

FARM COUPLE songs (tracks 1, 5, 17, and 18) most clearly, but the others fit pretty well.

ALLEGORIES (tracks 2, 3, 8, 11, 17). Aye; more allegorical than usual.

Who are ‘the farm couple’ then? She is a grown woman on earth variously known as “Lola”; a mad horse; Birch’s mother; Dick Turpin; a Nevadan; Esme’s adorer; “slow-heart”; Joanna Newsom. I’ll call her J.

He is, variously: “King Ludwig I”; “dragon”; “Bess”; Bluebeard; a magpie and a bluejay; a wolf-spider; a “silly goose”; “long-life”; and various hinted-at male celebrities whom I’m not interested in gaping at. Call him B.

We’ve only clues. I say “Newsom” when I mean “the songwriter”, and “J” for the protagonist - nothing more presumptuous (history is just organised gossip).

I don’t believe in overreading. Interpretations are second-order features, and if you honestly see [x] in a thing, then [x] is there.

The album has an arc: from the courting of “Easy” to the final moving out of the shared apartment in “It Does Not Suffice”. Each disc has its own subarc

too (consider the mood swing between “Esme” and the next track), but I’m less clear on those.

---

### Disc 1

```
{%      include hoom/easy.html      %}
{%      include hoom/hoom.html      %}
{%      include hoom/81.html        %}
{%      include hoom/good.html      %}
{%      include hoom/noprov.md      %}
{%      include hoom/baby.md        %}
```

---

### Disc 2

```
{%      include hoom/ona.html      %}
{%      include hoom/bess.html      %}
{%      include hoom/cali.html      %}
{%      include hoom/jack.html      %}
{%      include hoom/go.html        %}
{%      include hoom/occid.md        %}
```

---

### Disc 3

```
{%      include hoom/chalk.html      %}
{%      include hoom/esme.html      %}
{%      include hoom/autumn.html     %}
{%      include hoom/ribbon.html     %}
{%      include hoom/king.html       %}
{%      include hoom/suffice.html    %}
```

---

### Misc notes

```
{%      include hoom/voice.html      %}
<h3>DRUM</h3>
<div>
    Neal Morgan's percussion work (writing and performing) is the only voice in the thing th
    He's got a particular jazz-born genius, but it's more obvious live. He makes the first o
</div>
{%- include hoom/foots.html %} {%- include lazyload.html %}
```

# Staring at mystics

Gavin

2011-01-02

```
{% assign ana = "https://en.wikipedia.org/wiki/Analytic_philosophy#Analytic_metaphysics"
%}
<h3>2020 update</h3>
<div>
  One of the odder rabbit holes I ever went down:<br><br>
  I was so contrarian as a young man that I spent an entire month reading writers I myself
</div>
```

(c) Roger Penrose, 1999

Can it be that so many men, of various times and nations, outstanding minds among them, have devoted so much effort, and indeed fervor, to metaphysics, when this consists of nothing more than words strung together without sense?

– Rudolf Carnap

I think now that the right thing to do would be to begin my book with remarks about metaphysics as a kind of magic. But in doing this I must neither speak in defence of magic nor ridicule it. In this context, in fact, excluding magic has the character of magic.

– Ludwig Wittgenstein, PI manuscript

[Mysticism is] a philosophical urge gone wrong. Thousands of lesser philosophers are always with us to prove that it can go more wrong still, by trying to form systems out of no knowledge at all... The occult and mystical are perennial short cuts to a supervening vision... it is quite possible for the subtle visionary and the shouting dunce to inhabit the same skull... the essential truth about people prone to catch-all theories is that they aren't in search of the truth, they're in search of themselves.

– Clive James

People don't read philosophy. What do the pathologically open-minded people of the world read instead?

Bookshops tend to have only one shelf of philosophy, if that - and eight of something they call "*Mind, Body and Spirit*": books trafficking in sentimental,

pseudo-philosophical, pseudo-psychological superstition-porn. One step removed from Scientology.

## Why care?

The most popular philosophers in the world do not receive any professional attention: they're beneath notice. I call them the philosophical underworld. I refuse to dismiss them all just to fit in. Further: never mind **true; these ideas are loved**. They are the livelihood of four generations of global subculture. No matter how ill-founded, ill-grounded or even actively destructive, this gives studying them value.

```
{% include mystic/caveats.html %}
```

We say there's "hard" subjects and "soft" subjects, with ductility proportional to mathematical rigour. Is there such a thing as "hard" mysticism?

Famously, mysticism came back in the 60s. It recurred for a number of reasons - a plausible sketch being that postwar disillusionment with the Western script of disenchanted materialism, led to consequent bad readings of Buddhism and Hinduism, the challenge of authority in art, romance, and war, fear of Cold War realities, sex freedom. But a minor reason is because the emergent Analytic philosophy, through its boring technical topics and formalism, withdrew from the public sphere. Rightly or wrongly, philosophy is considered the source of existential insights, and when it fails to supply it, less rational forces will supply.

There does not have to be intellectual dishonesty in holding that *there is more to this than this*. We've gotten used to branding this kind of thing 'mysticism'. So be it; but cut away the liars, Messiahs, irrationals and fanatics, leave in some of the schizophrenics, and you are left with the philosophical mystics.

Spinoza is the paradigm hard mystic. (This adds to the exemplar function he already serves for the groups "mechanical philosophers", "early liberals", "deductive rationalists" and "inspiring heretics".)

---

```
{% include mystic/list.html %}
```

It'd be a mistake to go to these people looking for The Answer. But this isn't what I go to Epicurus, Kant, or Dawkins for, either.

## See also

- Shalizi, Philosophies' Evil Twins

# ‘Why study economics?’

Gavin

2011-02-11

Yet good, or even competent, economists are the rarest of birds. An easy subject, at which very few excel! The paradox finds its explanation, perhaps, in that the master-economist must possess a rare combination of gifts. He must reach a high standard in several different directions and must combine talents not often found together. He must be mathematician, historian, statesman, philosopher — in some degree. He must understand symbols and speak in words...

– Keynes, eulogizing Alfred Marshall (and himself)

The question is two questions with a covert moral element: “why economics?”, “why university economics?”, and “why *should* one..?” abbreviated away.

## I – Why economics?

I study economics because

- I am angry, and I am no longer young enough to be angry without it. Part of my petulance comes from the fact that the world is not as good as it could be; the rest from there being no easy answers to the first fact. I want neither the self-exalting cynicism of conservatives nor the vague rage of anti-globalizers. Good economics is about challenging bad solutions from whoever pushes them.
- What rhetoric was to the Romans, economics is for us: that is, the favourite lever of the politician. Right or wrong, the economic mindset rules the world. Thou shalt incur cost only for benefit; thou shalt understand preferences as rational and transitive and existence as a function. “Policy” – which is always economic, if only because it has a budget – is where many of the largest things are done. And large things need doing.
- An illiterate society cannot really be democratic, and because what it means to be “literate” shifts, and should shift beyond just “recognising written words” to include something like “grasping the determining forces of the world, what we’re up to”. A great part of that would be economic literacy. Our intuitions regarding much of the causal history of macroeconomics are flatly wrong; so, even given neoclassical theory’s raft of assumptions, the framework at least generates educated guesses.

- As Keynes points out, the field is a chimera - bit of this, bit of maths, bit of overlooked politics. Life is a cross-breed too.

## **II – Why at university?**

Because thought happens inside this professional structure now, in professional notation, with a professional bearing. If I wish to think for a living, I have to do it to measure. At least university curricula are vetted and compressed: unlike some other fields, academic economists are quite often practitioners, and thus have a chance to know which techniques are actually used and what is frill. Then there's the culture of the profession, which is hard to get from outside.

As well as studying economies and people's economic behaviour, I want to study economics. A lazy accusation is that the field is obsessed with egoism, with rationalizing all behaviour as personal utility-maximisation. (This is lazy because it confuses one elementary model, the neoclassical consumer, with the received view of the field.) Actually, in practice it treats everyone as essentially unprincipled, which is not the same as egoistic. People aren't stupid, unless you count greed (maximising externalized preference) and laziness (hyperbolic time-discounting, the lack of optimisation) as stupidity.

I haven't heard the word "capitalism" in perhaps two hundred hours of lectures. But I don't imagine that physicists spend much time talking about inductive reasoning, nor theologians much on how they have to prove there's this 'God' figure; the market is integral to the particular theories we have built, and incentive analysis works for any system whatsoever. But a good capitalism needs humanistic economics.

Many economic studies are outright rude (e.g. revealed preferences regarding self-obsession, prostitution, cheating), and wound society's unearned sense of dignity. Academic status is a good mask to wear when uncovering these!

### **Shoulds**

Returning to the hidden 'should' in the question: the unwitting omission of the moral element is symbolic. Economics proceeds on the basis that its status as a science disentangles it from moral concerns. (Economics was once '*Political Economy*'. What we know as "incentives" were once "temptations".)

We look at the intersection of desire and environment. Both of these things are vast and vastly unstable, so it's only to be expected that the joint distribution of the two is even worse. Economics predicts - on average rightly - that the question people ask is: "What's in it for me?" The good economist replies, "A better world."

# ‘Infinite Jest’ (1996)

Gavin

2011-04-03

{% include jest/links.md %}

(c) Cody Hoyt (2009)

4 Dating App Profile Cliches You Can Ignore if He Has a Dog . . . Anything mentioning “Infinite Jest”. Yuck! I mean, besides the fact that David Foster Wallace was an abusive shitbag, a Tinder guy thinking he’s deep for having read a book that thousands of his contemporary pseudo-intellectual bros have also read is a red flag. But honestly, you’re absolutely allowed to ignore this one if he brings his chunky lil’ english bulldog on your coffee date.

– Reductress

People like David Foster Wallace. But pretentious people like him, because his big book is difficult and they think his status will rub off on them; and but he did some horrible things, and so this large book’s reputation is much larger still than it (e.g. As well as the usual exhaustive wiki cult attentions, there’s a series of wacky blogs and a support group devoted to how gruelling it is; we view length as pretentious in itself, which speaks badly of our motives or attention span), and so you have to begin your discussion of this beautiful, tragic, *silly* thing clarifying that you’re not like those other guys. Or maybe you can gesture towards doing that and then say no, I’m not going to do a disclaimer, if I don’t have to do one for liking fucking Hamsun or Celine or London or Dahl or Althusser or Mailer or Koestler or Lakatos or Angela Davis or AA Gill or Malcolm X or Alice Sheldon or Stein or Burroughs or Pound’s writing.

The older Mario gets, the more confused he gets about the fact that everyone at E.T.A. over the age of [10] finds stuff that is really real uncomfortable and they get embarrassed. It’s like there’s some rule that real stuff can only get mentioned if everybody rolls their eyes or laughs in a way that isn’t happy.

Despite appearances, it has a very simple message. It’s about the very real downside to being (hyper) intellectual: that your theories can get in the way of your physical sensations (rob them of their immediacy and emotional impact) and prevent you from interacting with others in an easy, fluent, direct way. It’s about the rejection of postmodernism from within. But these are pretty niche disorders. Much more importantly it’s about (what Wallace saw as) the

general late-C20th tendency towards a toxic sort of irony, which destroys value by making people less receptive to the world, and which emotionally dulls those who take contingency to heart.

This is now called “the meaning crisis” and there’s a large collection of internet people talking about it as if it was the most important problem in the world. I don’t know to what extent our crisis is the same as DFW’s one; I don’t know to what extent this is a problem for one sort of sceptical Western intellectual and no one else. Insofar as you think Nietzsche predicted it correctly in 1880, it might be the same and a general problem.

But *Infinite Jest* distracts you from those simple meanings with a forest of calculus and psychopharmacology and Boston slang, with 200 footnote discontinuities and 7000 neologisms and proper nouns. I say that Wallace “rejects postmodernism” - whatever that means - but he stubbornly maintains the confusing, excessive, perspectival, mashup aesthetics which are the least fake denotation of the term. (In the last 20 years people have painstakingly built tools to clean the mess intentionally strewn before you.)

This message is essentially the same as a thousand Sunday sermons: “*be excellent to each other*”, “*caring is cool*”, “*only connect*”. If it was not wrapped in armour-piercing arcana, fancy theory, and formal experimentation then its intended audience would never let it in. *Infinite Jest* has to be pretentious, because its audience is.

---

...I am just about the world’s worst source of info on Infinite Jest.

- Wallace, letter to fan

It’s hard to say things about *IJ* because, despite the above quote, in a real sense you are *competing* with Wallace if you do; *IJ* has already Freuded, Hegelled and problematized itself, not least in its 200pp of (plot-endogenous) footnotes. It also has no ending: you shlep through a thousand pages, work for weeks, and are rewarded with a slap.

It reports neurotic details of a dozen things I’m not interested in - tennis, optical physics, pharmacology, counter-pharmacology, the specifics of child abuse - and is riveting even then. Every hundred pages there’s a passage to gasp and half-close your eyes at. It is warmth reporting on ice.

---

### Misc notes

- James is Wallace; the samizdat is *IJ*. Both are over-the-top, both are missiles aimed at emotional detachment.
- Hal’s mental illness is overdetermined: we’re given half a dozen possible causes for his detachment from the world. Against the grand cultural point

of the book in general, these are biological: drug withdrawal, drug toxicity, an exotic mould he ate, (plus burnout from the strictures of elite athletic training).

- Above, I focussed on the personal emotional-philosophical stuff. But there are a dozen plot threads, including an apocalyptic terrorism plot, and an idiot celebrity germaphobe president.
- There will be overlaps with its namesake *Hamlet*, though not for me, not yet, barring: “BERNARDO: Who’s there?” - opening of *Hamlet* “I am...” - opening of *IJ* As well as the bit where James Incandenza (the father)’s ghost manifesting and warning... a character he’s not related to - who notes that had the ghost appeared to his son, he would’ve messed the kid up... As well as Hal spending much of the second half of the book doing nothing brooding.
- *IJ* stylizes itself with things which have been considered the *opposite* of style - formal organisation titles, straightfaced repetition of details and nerdy facts and full names; unnecessary, often-unfunny subject-predicate clarifications (Wallace, that is); and oodles of technical explanations. The thousand footnotes give reading it an interruptive rhythm. So but there’s constant digression in the text (at one point there’s three pages of flashback and tangents between two lines of dialogue) and in your train of thought. Life is a series of more or less successful digressions.
- Almost everyone is in some way deformed: phobic, neurotic, addicted, displacing, disabled. It would be easy to assume that this is Wallace’s view of us all, but *IJ* is cartoonish and deformed in a lot of ways.
- DFW is an omnivore, a generalist: *IJ* is nauseatingly detailed with academic arcana, medical/chemical/medical/mathematical/scientific passages, 1C20th Boston slang, film-geek waffle, & what one reviewer called “pseudo-science” (but which are surely just “alt.hypotheses”) - which theoretics all add up to sensory overload, and exasperation for anyone who expects to encircle and dominate what they read with their understanding.
- The “unreliable narrator” conceit in literature is making its worthy way towards cliché; the third-person-objective narrator who is nonetheless occasionally *ignorant* is entrenched but still crisp - but ignorant *footnotes*?
- The discourse changes style and inflection when swapping storyline to storyline - most noticeably when the Francophone Marathe is its object. (At one point I got suitably paranoid and saw the whole book as an informal report by the cross-dressing secret agent Steeply.)
- The physical contrast between brothers (Apollonian, Olympean) Hal and (Tiny Tim, deformed, innocent) Mario is unsubtle, but so. Mario and Lyle are perhaps the only naive, unironizing characters among, say, the hundred in the cast. This links Mario’s innocence to his defect: innocence is a

“defect” in an ironic world. And ‘stupidity as innocence’, too: stupidity as the absence of an attitude, rather than the absence of intelligence.

- Like Don DeLillo or Orson Scott Card, Wallace makes his children ridiculously hyperarticulate. I’m inclined to name this sort of thing “Hogwarts Syndrome”, with the kids more sensible, prolix and interesting than any pack of children have rights to be.
- Mario notes at one point that he has lost his easy empathy with his little brother, that he cannot tell how Hal is feeling anymore: we the readers go through the same, beginning the book inside Hal’s head at a moment of trauma and insight, and but gradually (as the cast expands) lose this closeness.
- The word “annular” recurs every thirty pages, though I only noticed this cause I had no idea what it meant. (“*... of or pertaining to a ring or rings, ring-formed, ringed.*”) I now think it’s a key MacGuffin, describing as it does
  - how *IJ*’s cold fusion works;
  - how (super-MacGuffin) James Incandenza’s film oeuvre is structured;
  - how addiction works;
  - the appeal of suicide;
  - how they cured cancer by giving cancer cancer;
  - maybe the “Subsidized”, ruined nature of time in his near-future paratopia;
  - and *IJ* itself - how its storylines fit (rings-within-*IJ*’s-ring). He could have used “meta-”. It wasn’t ruined in ’96.

---

### Is irony toxic?

The topic of futility would arise only if one were trying to surmount time, chance, and self-description by discovering something more powerful than any of these. For Proust and Nietzsche, however, there is nothing more powerful or important than self-redescription. They are not trying to surmount time and chance, but to use them. They are quite aware that what counts as resolution, perfection, and autonomy will always be a function of when one happens to die or to go mad. But this relativity does not entail futility. For there is no big secret which the ironist hopes to discover, and which he might die or decay before discovering. There are only little mortal things to be rearranged by being redescribed.

– Rorty

The other great clear postmodernist, Richard Rorty, actually celebrates irony (though it’s not quite the same thing that Wallace is attacking). Irony is like (radical) scepticism plus the pragmatic sense that you have to take *some* stance towards the world. So you admit that e.g. human rights are a Eurocentric

construction, that you affirm them entirely due to an accident of birth and history, *but you still insist on them*.

It's a philosophical question whether this makes any sense, whether it is actually impossible to obtain moral truths, whether intercultural comparison is valid. It's an empirical question whether any human can be happy not taking things for granted, admitting that their worldview is arbitrary.

---

### Term

There are six suicides in the book, not counting people who watch the *samizdat*. Joelle, Gompert, Day give long rationales, among others (eg. p648):

the person in whom Its invisible agony reaches a certain unendurable level will kill herself the same way a trapped person will eventually jump from the window of a burning high-rise. It is the weighing of two terrors, a rational decision, which rationality is invisible until you are there with the flames at your back...

This can't help but resonate now. Just because you're a genius doesn't mean you'll ever arrive at any answers.

---

That a book about the importance of sincerity became, first, the object of a cheap signalling game and, subsequently, the object of scorn and the received epitome of pretension, is just one of those fucking things.

---

### See also

- Against the Culture
- Aaron Swartz, who had a similar disposition.

# A Sentimental Journey Through Parts of England

Gavin

2011-08-17



Allons! to that which is endless, as it was beginningless, To undergo much, tramps of days, rests of nights, To merge all in the travel they tend to, and the days and nights they tend to, Again to merge them in the start of superior journeys...

- Walt Whitman

It is an age so full of light, that there is scarce a country or corner in Europe whose beams are not crossed and interchanged with others. - Knowledge in most of its branches, and in most affairs, is like music in an Italian street, whereof those may partake who pay nothing. - But there is no nation under heaven abounding with more variety of learning, where the sciences may be more fitly woo'd, or more surely won, than here, - where art is encouraged, and will so soon rise high, - where Nature (take her altogether) has so little to answer for, - and, to close all, where there is more wit and variety of character to feed the mind with: - Where then, my dear countrymen, are you going?

- Laurence Sterne, *A Sentimental Journey*



## Day 1: Aberdeen to Carlisle

For simplicity I tell people this is my first time in England. The truth:

1991: Visit to Greater London for a wedding. 2 days.

1996: Visit to York for nothing in particular. 3 days.

1999: Visit to Windermere for my gran's Mormon nonsense. 3 days.

2002: Through it on a bus to France. <1 day.

2004: Through it on a bus to France. <1 day.

So actually close enough. Into Carlisle about 7pm. Struck by the sheer amount of brick, especially in confectionary red-white-red-white pattern. Panic rises when the rain comes on: am hungry, alone, and have no idea how to get out

of town. Rank my paranoias to pass time: fear of criminal folk (low); fear of sarcastic people (high); fear of illness from rain (med); fear for bike (med); fear for no campsite (critical). Eat in a truckstop, and meet George, a magnanimous gobshite who introduces himself as a connoisseur of rucksacks, and offers a campsite in his garden within 30 seconds of meeting: “Life’s too short to be horrible to people”. Sleep in a hedge on an industrial estate instead, for some reason. Possibly manners.



## Day 2: Carlisle to Patterdale

Give up trying to sleep at 4am and get on the A6. Leaving, I have Carlisle to myself, sharing only with the odd Stobart man. Scoot around the castle happily. Bike the 20mi to Penrith, get there by 7am, ffffuck. Do not feel remotely good - lurk in the sun, chilled and sick. Seem to eat and drink continuously. Struck by the casual fellowship of the unmotorised: hikers grimace at each other; cyclists always nod. Come upon the absurdly picturesque Patterdale, facing the unspeakably picturesque Ullswater. Every other house is a B&B. Climb two miles up a fucking mountain to reach the Arthurian, the sublime, the baffling YHA Helvellyn. It’s empty and unlocked, so I help the recycling man:

TONY: I were a livestockman f' the Earl f' 44 years. (“yeus”)I: What changed?  
TONY: Got bored.

Middle-aged hostel crowd are great - we all stink, I trust them instinctively. Shower, pathetic with gratitude - only 28 hours since my last. Giddy with fatigue, I entertain at the dinner table. Honestly not really sure what I was saying to them.

## Day 3: Helvellyn to Grasmere

Sit on a crop over the Ullswater all morning. Read Auden ('Paysage Moralise' impresses me into a deep funk). Meet more people - it's just that kind of place. Listen to Some Call It Godcore; perfect. Bike making an irritating noise - sit down to bodge it; it's miraculous that it was running at all; chain had pulled off the deraileur and was only going by cutting a groove into the guard. (Things fall apart, but sometimes they fall into place at the same time.) Up Birkbeck, harrrd. Feel great, have lunch at 1500 feet. Soon after, coming down "The Struggle" (too fast, too topheavy) crash pretty badly, using my skin as a brake.

If I should fall, think of this of me,  
That there's some corner of Cumbria road  
That is forever Gavin.

Man happens up the road a minute later and kindly offers first aid: through my teeth I fail to express how much this means to me. Limp to Ambleside, brakes on, then on a little further to Grasmere. Have earned my dopamine today. Inexplicably, the hostel has a copy of Jung's Archetypes, and some Sloterdijk in

Dutch(!) While wrestling with Jung, eavesdrop on loud Australian girl: "Efter Glestonbury, yea, eeverything was all so 'ohmygod toomanypeople' so we weent to the Laykes... and thain during Paul Simon oi felt raily seack..."

Does everyone's tourism seem contemptible but our own?

## Day 4: Grasmere to Kendal

"The loveliest spot that man hath ever found." - Wordsworth

Grasmere's nothing to Ullswater. (see how quickly we become worldly! J later mocks the fact that everybody he knows seems to profess a working knowledge of and affinity for the Lakes.) Sleep is difficult without elbowskin. Go see Dove Cottage and Rydal Mount. I actually gasp once let out of the former, bloody guided tour. Japanese people everywhere. I side with Hunt against Wordsworth: a poet who withdraws from humanity needs to justify herself, and WW has no grounds but beauty for his endless vacation. Through Ambleside, lunch at the Priest Hole. Struggle the 4 miles to Windermere, and would've done without the wounds screaming. Bugger about for a couple of hours, preparing to seem authoritative rather than bewildered (R and J are coming). He arrives eventually, and we go fail to find a 24hr shop. Camp by an empty house. Criteria:

Drainage?

Windbreak?

Soft ground?

Teen haunt?

Noise?

Legal?

Unsurprisingly, having a tent is an exponentially better way to be.

## Day 5: more Kendal.

Woken by a bemused builder at 9am. Neither of us knows what the other is doing there.

Day of forced grace - me and R waiting for J. Eat beans by the river, eat Mint Cake (CHRIST) by the Castle. Mint Cake is to meringue what diamond is to coal. It is good indeed to have a companion, but I feel the need to mask my leg-weakness and pain, even so. Jokes are real again! (Donald Trump's helicopter's tapeplayer has one Kraftwerk album jammed in it.) Visit a church (St Thomas'?) - grotesque.Sun.J is late (straight in from Zambia). Eat crap, then go see Half Man Half Biscuit. Hits-laden set, smelly and fun.

We camp desperately, end up next to a railtrack among thistles. Laugh.

## **Day 6: Kendal to Sedbergh**

Dunno how we slept - “Hitchcockian” night. Heat rises to about 28 celsius as we approach the Dales: and it’s 18 miles of uphill. J suffers. Reach Sedburgh, where we founder - J’s bike is fucked, and the nearest bike shop is Kendal, and you can’t take bikes on busses, and (...) Cook dhal in a churchyard, and laugh. J returns to Kendal. Bugger about with some philosophy history - who is the empiricist who connects Newton with Russell (if anyone)? - and go find somewhere to camp. First path we roll up, Ghyll Farm, agrees with grand nonchalance. My heart swells with the kindness, and the £100 view. Wash naked in a stream, good. James succeeds & returns taxied by Mark E Smith, good. Sunburn; ah whatthell. We eat peas in the pod, good.



## **Day 7: Sedbergh to Leyburn**

Passing a newsagents, note Guardian scoop about Milly Dowler’s phone with a theatrical “fffffuck!”. See a black rabbit. (So much roadkill in Yorkshire.) Begin the

Day begins with a road dispute, I solved. Road to Hawes is beautiful but painful. Seems that every rural Northern town is a “book town”. Eat total crap in a coldheart Hawes cafe. It’s too hot to work at 2, so we lie in a park. My backpack is too big, but there’s nothing I can throw out (only three changes of clothes, for instance). Press on to Leyburn, whose name no one can retain. Laugh. JW cancels our Middlesbrough appointment. We huzzah, a bit. Camp secretly at Stoop House Farm. (Sheep never shut the fuck up.)



## **Day 8: Leyburn to Ripon (Seven Bridges Valley)**

Nice downhill to the A6108. R ill (dehydrated). Heat punishing even at 10am. Spend a nice hour in the shade, texting while R sleeps and J scouts. Ripon, seeped in a seeping cathedral. I am learning to treat churches as a form of entertainment - a vital skill in the C20th, but yielding less these days. RC is good and squat. Light a candle (for gays killed by Christians).

Camp by Fountains Abbey, nervily. Make stirfry and inane entertainment - watching a wheel spin. Laugh... Brought Hume’s *Treatise* with me, but I haven’t read fuck-all all trip: The habit has been pumped out of me - and anyway the intro’s by some snotty ’50s Analytic no thankyou very much.



## **Day 9: 7BV to Thirsk**

Yorkshire is fucking huge, but its cultural footprint is very small, somehow. Have breakfast at the Fountains visitor centre. (Eggs benedict is a bloody weird dish.) Don't go in. Back to Ripon to sit out the Ferment (unbearable mid

Day heat). To Thirsk. S'ok - but the Blacksmith's Inn's jukebox, by this, mp3-dead, tune-starved point, is a revelation. With J, pump £5 in. O the fun of infliction. Out. Ask at a farm, who agree reluctantly. Owner looks like Harold Wilson and talks in a way that I adore but am crap at bantering with. Stove soy-bolognese and silence.

## **Day 10: Thirsk to Malton**

First rain during night. R's stuff damp but ok. Task #1: climb Sutton bloody Bank, a job that a bartender last night grinned maliciously at. As soon as we're up, first real rain starts. Soaked to my pants in 4 minutes flat. Hide in a visitor centre, but soon out Into It. Ugggh execrable. Biking fast and risky (13 miles in an hour). Feet three kilos heavier from waterlogging, visibility nought. R avuncular throughout. One facial expression for that hour:

>:X

Suddenly dries. Go to Castle Howard, brideshead visited. R and I pad around the ground barefoot and talk metahistory. J doesn't want to pay the entrance fee. Lots of pagan biz around. Fear the rain. Drink in Malton's "Crossed Keys", a warmly bizarre, Thai-themed, Medieval-catacombed pub. Tastes of doom, sadly. Laugh anyway. Eat mexican: yaas. J falls off his bike suddenly on the way out, minor but galling for all that. Then R's gear-cable snaps. Thus hunt for shelter early. Find another indifferent farm - J is the first to identify its "wrongness". The doom of canyons. Shave my lips and furrow my brow.



## **Day 11: Malton to York**

Wind torments us all night.

"J: How many stains do you have on you?" I: Uhh... J: I've got blood, oil, nutella, paint, suncream, bolognese, pen, grass, savlon and toothpaste on these trousers."

Eat nutella and leave with haste. "Easy" route is hard, uncertain and hit by crosswinds. Tempers fray on all parts. Wash my feet in a river while J goes begging for water and R frowns at the horizon. Called a gaylord on the back road to York (by a stranger that is, not J). Starving, disproportionately weary, and then I drink some off milk. (Veganism is impossible on the road.) Make it though, and eat in a bistro. Doze. Get a fusty 3-bed room. Decor: avocado plastic and forty-year-old taupe. Since I'm a vegetarian, R and J are designated

“normals”. Go see Potiche and struggle to find late night food again. York is good tho.



### **Day 12: more York.**

Absymal night - feverish, neuralgic, tinnitic, insomniac, gut-rotten. Sleep naked & still boil in my bag. It breaks around 5am, and I collapse. Shitting blood and farting butane. Has been coming for a while, now I think about it. With no destination today, I have time to break down. Breakfast is served by B&B woman’s children, ew. Saunter.J spots a woman having a ‘Proustian moment’ at a sweetshop window. She’s transfixed, mouth slightly open, eyes glazed. We stand a little while watching, before shame overtakes me and I wheel away. J protests that he thought it beautiful but: even so. National Railway Museum is impressive - full of things built for incredible wear, so you can poke and touch whatever. Face off against the most intrinsically evil train in the world. It’s while spiritually wrestling before it I work out that I can stop the pain if I don’t move and don’t breathe. Circle York some more - the conversation unwittingly(?) centring on our futures. Eat in unabashedly hippy veggie restaurant. (Meh.) To the Minster for chorales: a deadening sort of awe. Saw a bouncer earlier who was a human crow - jerky, wary and cruel. The Gabrieli Consort are human eagles (superlative) but also kiwi-birds (full of something larger than themselves). Afterwards, more blood.



### **Day 13: York to Harrogate**

Pain largely lifts. Change a tyre, read aunt Guardian, and away. Pain returns from exertion. Lunch in Weatherby, where argument about Class vis-a-vis delicatessens kicks off. Also re: ciabatta. A dull town. To Harrogate, full of parks. (To do list: feel good.) Eat bad masala in a park and laugh. Go see Lady In the Van. Theatre is airless and womb-hot. Play’s beautiful tho - the soul in question offering a lesson after all. Sleepy, I mistake the interval for the end. Rush out, find a farm, sleep in the calving field. Set up in the dark.



### **Day 14: Harrogate into Bronteland.**

Self-righteous passerby asks if we have permission to camp: his transformation after being rebuffed, like a balloon farting flat. Just after we decamp, the Rain comes on. A group hysteria comes on too - we dump R’s tent, flee back down the hill. Sodden breakfast in Cafe Rouge. More laughter. J leaving on a wet train. Blunder out of town and do 25 miles in 3 hours. Dry in the wind. The Bradford valley is amazing, Italian. Stop in Saltaire for ‘lunch’ (see photo).To

Bingley, which we soon retreat from, set up on a grand piece on nasty scrubland. I walk a mile and back to buy water. Talk metaphilosophy. Dream about home.



### **Day 15: Bingley to Littleborough**

Wake to find a passing dog has eaten my breakfast. Eat cereal on the verge of an A-road. We climb 1500 feet in two miles - on the edge of the Pennines now. Haworth is bloody dramatic - all 40 degree valleys and Burtonesque outcrops. More bland road towns, and then O! A highlight of the whole trip: an incredible, four-mile-long regular downslope to Hebden Bridge. Didn't pedal once. Shower at Todmorden and doze on an ex-golfcourse (what a lovely concept that is!) Power on to Summit Quarry, a stunning but midgey campsite. Tomato is the travel staple for some reason: 5/6 of the meals centre on it.

### **Day 16: Littleborough to Knutsford**

Sheep creep deep as we sleep. It's not pleasant going in these parts - though the sun's not intense, we have to take big roads. Decide to skip urbania - but Oldham train station is no more, so first we look for somewhere not grim to eat (fail, so first Wetherspoon's). Lots of hassle on road. Go to Ashton.

It takes us three hours to make our connecting train through Manc to Knutsford. R goes for food, takes 50 minutes at it, which crosses my Gerald Horizon (the time waited after the expected return-time of someone before you assume that massive disaster has befallen them). We are identified by one of J's lovely mates. Sleep in a barn. FTW

### **Day 17: Knutsford to Helsby**

Don't want to get up. Barns be comfy. Sit on a bench in Northwich and watch strange strangers. A Securicor man stops to pet a dog on his way in to Tesco. Stops for an unseemly length of time, really. Helsby is unsignposted. We muddle on. Somewhat frayed again. Arrive at J's house after hours. Shower, eat, miss a window for a Hong Kong internship, poke around. I am three steps into his house before I see a Burial album, to be fair. Up Helsby Hill; it's a chemical, Lemon-Jelly land. J is I suppose only standardly impassive to his parents. Eat cornucopic curry. J drives us to Ellesmere to see Tree of Life, a messy and chewy old thing. Back, we sit around, read existentialists.



### **Day 18: Chester**

Thought process on the cycle route from H to Chester:

*relationship between capitalism and love*

*1: growth and industrialisation invented the concept and supply of “leisure time” - a vital component in gardening one’s romantic love.*

*2: Concentrated mass housing allows couples to separate without the fear of homelessness-or-exile holding them together.*

Town's reet nice. J is unerring and charges around his teen haunt in a funny laconic way, but we blunder into the sights anyway: buskers, the weird double-shops, the Cryer, the Cathedral, a Roman, the walls. Refectory is good. Tension flares somewhat over my nonexistent road skillz.

Poem for Chester's shot-tower:

and I, a liquid falling, and morphing (as you do) and pausing on the water sphering anew.

Good hard towels at J's. More curry, then rush comically around Cheshire looking for J's friends in the old manner, blind guesswork. Nonbonfire party at friend-of-a-friend-of-a-friend Howard's. Thus get to see J in his hatching habitat, which is very good fun:

“Buuhhhhhh, mi nam is J. H., I do nut kno bout art or fukkn books.”

“This guy had shit himself in the club. But he were so cool about it, we felt like idiots for not shitting ourselves!” “When in Rome, shit in the woods.”

He also gets piled on, in the course of which I clonk heads badly with another of his lovely mates. I'm totally fine, but he got fucked up. Giant swollen eye. Helplessly embarrassed.

Banter is fast and riotous; I don't keep up. He is forced to be young here. It might grieve me that I exert little of this fresh pressure, but nah SUCH LADS

## Day 19: Liverpool

Another deceptively aimless one.

Day out (aimed well by J). Drives us at length through bewildering spaghetti roads. Knows his history, especially the Beatles tour less travelled. First stop, the Metropolitan Cathedral. It is virtuoso, powerfully unsentimental. Dozens of excellent moments are set into a shocking overarching theme. It is my favourite church. I know this simply cos for once I felt no contempt in it. To ‘the Phil’, the Docks and the Tate. It's Magritte at the moment, and he is loads of fun. Also the other Cathedral: equally brutalist but less modern, less true. They face each other down Hope Street, and the smaller Metro guts the shit out of him.

```



```

Back, risotto trop vert, and to the pub. Meet Cheshtronica supremo TNJX, very shy - well. Moments of pure fun - “Sad or Bad?!” and seven people shouting at a quiz machine. J drives well, well drunk.

## Day 20: Manchester

“I can suck melancholy from song as a weasel sucks eggs!” - Jacques

Roll straight out of bed onto the 1100 train to Manchester. No conductor! Rain forces us into weird cultural junkshop. Buzz about - go to the gay bit for Turing, the crap goth-encrusted cathedral, the cathedralite Town Hall. Try to see Johnny Vegas, but are crushed. Always so much about soldiers in churches. It’s enough to drive you Marxist. Manc Cath claims all of one regiment’s late-C20th dead on its wall - though a decent number of them wouldn’t have given a shit about Christ. Supernatural insurance has done this world too much harm. Big public gallery is oddly dissatisfying, but there’s a good Turner and some fun contemps. Try to eat in the modish Northern Quarter - no seats in: “Common”, “Trof”, “O??”, nor “Oklahoma”, so we eat in a mediocre vegan place. See Craig Charles.

See *As You Like It*, proper good even in the Gods. Another free train back, chatting with a friendly drunk in a borrowed suit.

## Day 21: the Wirral

One last sally. To Jodrell Bank, where we stand for 5 minutes and leave. Am quiet, fatigued in some occluded way. J puts on Beastie Boys and Half Man Half Biscuit in the car; I could just stay in all day. Do plaques for two friends, Peel and Blackwell. Rain is filthy thick. Have a “euphorically” bad time in Birkenhead. Eat in Wetherspoon’s, inevitably. Fail to get a present for J’s parents. Internet for a bit, trying to fling myself into a fruitful future. (A press pass to the Edinburgh Festival, and the groundwork for Low Lands, my book on nationality.) Go a to pub quiz, at which J is extremely unhappy.

“Which Canada-born Bayern Munich midfielder made his international debut against the Netherlands in 2001?”: “WHAT THE FUCK!!!!!!!!!!!!!!!”.

We are not good but are funny.

Moment of poetry in a grim place (toilets of the Belle Monte pub, Frodsham): remarked to J how lovely his friend P is -

J: He is, isn’t he? Famously is. He thinks I hate him though. I: [grandiose] So let him know!

Later learn that P was in the cubicle during this, trying to save our modesty by calling out, but was ignored.

## **Day 22: Helsby to Congleton**

Woken by J having a good idea in the next room. Onwards. (J stays behind.) Back immediately with some nice STEALTH WINE for hosts. To Nantwich quickly. Cheerfully piece together a lunch of chips and redcurrants. Staffordshire is fucking nasty. Not much thinking or talking involved, just grit teeth and get through it. Make it to Congleton about 7pm, have a drink. Meet our first Great English Eccentric, a rude hag living in a giant scrapheap of fiftyyearold sports cars. Camp in a swamp on top of a hill. DONE.(Concept without a word: beaux vivant; informal artist; self-artist. Someone who lives according to taste, and unnoticed acts of art. "Aesthete" comes close. What would Nietzsche call them?)



## **Day 23: Congleton to Ashbourne**

Slept about 12 hours in tolerable misery. Spend ages decamping, soggy and grumpy. Less dashing, less swashing. Uphill struggle to Leek for hours - a spry 60 yearold biker overtakes us. Go to a crap cafe who begrudge us a sale. Limp on, fringing the Peaks.

Altered consciousness, really. Focus on ruining your legs. Did about 40 mile yesterday; apparently that was too much. Stop at Ashbourne cemetary (count on the dead for peace). There's a humanising flash of sunlight, but that's all. Drink and chat amidst bunting and Scottish flags (?) Thai restaurant for tea which had the same gorgeous dense sculpted tofu as had in Beijing. Scratch out another mile, camp illegally on a footpath.

birdsong as gunfire(which it's closer to than serenade)eternal woodland carnage,interminable grudge,the cock a flare up-arching, and the warring won't be budged: wings like old m.g blues.

## **Day 24: Ashbourne to Smisby**

Hypersensitive night - too close to the road, too much stink. Another supermarket-bench breakfast and away. Fatigue lifts. Into Derby on a wave of admiration: bike paths everywhere, big news screen and a stylish contempt for its past. Shower, do cathedral, do pub. It occurs to me that EngSoc are the natural enemies of Philsoc. It also occurs to me that I want to set up a intersociety football tournament. Moment of sublime error when I leave my bike-lock key somewhere. Set the staff of Wetherspoon's searching for it before noticing it myself under the table. We leave hastily.Except you can't, because the outskirts last 10 bloody miles. Derby hath spread her wings. Along the Swarkstone Causeway, which is a stunning thing to leave in the middle of nowhere. We're almost efficient. Face the first real campsite difficulty - ask at half a dozen houses, nothing. A nice man owns a wee unkempt field, though, and there we are.

## **Day 25: Smisby to Coventry**

People have been good to us on this trip. The only explanation they usually give is “No reason not to, is there?” or “Life’s too short to be unkind ter folk.” J mocked this when I put this to him, but the point is not that this is some metaethical epiphany, but that tacit proto-principles can support action on their own, and do, and maybe always have. Back road to Birmingham goes inexplicably easily; daydream all day, replaying the year. Stop in Atherstone and consider the lilies. Make hellish trip to Coleshill train station, which is 5 miles out of Coleshill through industrial hell. We stop in Birmingham for literally ten minutes and go to Coventry instead. Everything’s closing as we arrive. Eat Cantonese, enduring the worst that Cantopop can offer. After much blundering, find a weird empty grassland and camp. Toss, turn.

## **Day 26: Coventry to Callow**

Up in sunshine. R’s away today, so I trade my lovely inflatable mattress with him in exchange for the tent. Error expected. To Coventry Cathedral, lock our bikes beneath Satan.



We don’t stay long - there’s a Christian rock band practicing in the main cavern and I’d be claustrophobic even without the aural cack. Get overwhelmed - there’s a Hiroshima exhibit in a side-chapel, and something in me just gives way. I am about as fit as i’m ever going to get; I’ve got no physical complaints; I’ve done something grand with my July - but I’m not in good shape, in some important and wordless sense.

R away home. I’m unaccompanied for the first time in three weeks; overtones of fear arrive. Push out defiantly for Warwick Uni, a great glass dump of knowledge. I could deal with this for a year; 2013, say. The economists are in the same building and floor as the philosophers, but of course they never speak.

To Worcester by dark, whereupon I’m knackered. Their gallery’s crap. Eat trendily (Slug & Lettuce), shower at the leisure-centre with the broken boiler. Rain comes on so I ride around the empty ‘Crowngate’, their covered marketplace-mall. Well fun. Waves of despair come on - have to bike 4 miles out of town, get a dozen rejections in that space. Find some Commons and collapse. Sleeping on the actual ground is quite a skill - have to work your back into and around all the bumps. Or go mad.

## **Day 27: Callow End to Stow**

Woken by that most powerful of freejazz quartets, dogs screeching in joy near the ear. Even given that I’ve an awful unfocussed and heavy feeling, in this, our fifth week of the Road. Get a flurry of texts informing me of the passing of Amy Winehouse. (What’s the obscurest celebrity you’d text someone over?)

Make it to Great Malvern running on empty - no water nor nothin. Sit and eat for an hour.

Overheard: [patronising] "Darling, you are what you are - it's your genes, isn't it?"

On this spot in AD1211: "Darlinge, you are whatt you are - it be Providence, see you not?"

Nice town, full of theatres and Cryers. Up, up, up to Colwall, which is unspeakably peaceful. Lurk around the Downs School reading Auden. Then away, down to Stow. All the little towns around here are obviously nice, but I can't stand them. Spot something bizarre on a hill. We stare at each other for a bit until I give in. Wrestle over whether to get a hostel or eat well (both about £16, see). Hostel it is. Can't decide what I want first - a shower, some internet, clean shorts, food, wine, a nap, safety, a giant-ass bowl of cereals, or a little tenderness. Sort out my life; find all these things. My relief is violent.

### **Day 28: Stow to Oxford**

Bad breakfast and back to bed. Read the hostel's *Female Eunuch* all morning ( :o )

Road is easy, or perhaps I'm finally Road Worthy. Into Oxfordshire without breaking stride. Chipping Norton - the dead heart of the media-political complex - shows none of the emotions expected. Pass Blenheim Palace, the giant-ass pile we gave to Marlborough after he killed all those people for us at Blindheim, Bavaria. Doomed to live/drown in one

Day of your life. If the man had any sensitivity he'd have fucken ached. Oxford is crowded, loud, expensive and discordant. "Honey made stone" - no, mead-vomit frozen in place. Tbh I return the impression, being a thoroughly bad tourist - I rush around, frown, steal, piss on Magdalen College and leave. Almost lose my phone - leave it on a bench - which I do feel as a rebuke by the souls of this bloody place. Camp on a farm by Didcot. Farmer banters, but I'm having none of it by this stage (to my later shame). (Want to Stop, but honour dictates.)

### **Day 29: Oxford to Andover**

Heatstress busyness psychosis, eh? Hyperthermic tensomatic kinetic batshit, no? Stop over in Newbury, which is horseracing and nothingness. Stock reply to crap funeral: "She really loved language." Stock reply to crap beer: "You can really taste the hops." Stock reply to crap reply: "That would be an ecumenical matter." An important thing to know about yourself: you have a limited appetite even for beauty and novelty and adventure. Around three weeks, cynicism and impermeability of the soul begin their encroach. I don't find a hostel. The passion with which I want to get on with my life would make Nietzsche proud.

## Day 30: Andover to Salisbury

Sick of the same sweat-caked clothes. Sick of a sore back. Sick of shit beard. Sick of focussing on negatives. The ‘Plain’ is impressive. Ride around the cathedral a few times looking for Jonathan Meades’ childhood house (failure). Go down “Endless Street” just to be contrary. YHA Salisbury is pleasingly shambolic. Sit on a picnic bench outside, drink three beers, eat an entire pizza and watch the sunset. Pangs for someone to share the moment with. Think about prankster philosophers. (As we go, the comic portion of the work increases)

Diogenes -> Montaigne -> Nietzsche -> Derrida -> Zizek.



## Day 31: Salisbury to Elgin

Dorm was incredibly silent for a room of eight men sleeping. Not sure when I decided that Salisbury was the new terminus, but the thought of more south makes me angry by this stage. Go to the Cathedral, and then fuck off out of it.

Book I got at an Oxfam is fairly amazin:

<blockquote>I beseech you! if ever we shared philosophical impulses, take responsibility for</blockquote>

- Jaspers to Heidegger

(Wish I could have as dramatic a conversation about this abstract a thing with as dear a friend of mine. Puh. Stupid, pragmatic, anti-Nazi friends.)

The train undoes a month’s work in a quarter of a day. As is my new habit, take the train to Elgin instead of the bus. Use my Aberdeen ticket to get to Elgin, somehow (conductor possibly saw my expression). Elated, trippy journey home in the dark from Elgin - there’s no streetlights on the main road, so every time a car approaches I’m completely blinded. Heidegger in my forebrain and epoché on my mind.

What should they know of Scotland who only Scotland know?

# Rubinations

Gavin

2011-10-20

```
{% assign rubin = "https://en.wikipedia.org/wiki/Rick_Rubin" %} {% assign
playlist = "https://open.spotify.com/playlist/6NyWwyOC84YiTkaTklMtQj" %}
{% assign rose = "http://www.metacritic.com/music/van-lear-rose" %} {% assign
stap = "https://en.wikipedia.org/wiki/You_Are_Not_Alone_(Mavis_Staples_album)" %}
{% assign hard = "http://en.wikipedia.org/wiki/Hard_Again" %} {% assign
ldn = "http://en.wikipedia.org/wiki/Howlin%27_Wolf_London_Sessions" %}
```

## Playlist

One of the nicer things in the last 20 years of pop music are rubinations, after Rick Rubin, an early innovator in it:

an over-the-hill musician

is renewed, accrues critical acclaim

from working with a young svengali producer,

on an album containing covers (especially surprising ones).

The festival circuit

or very large sales follow.

May-September music.

- Johnny Cash (& Rick Rubin) - American Recordings (1994-2003). Satisfies #1, 2, 3, 4, 6.
- Loretta Lynn (& Jack White) - on 2004's Van Lear Rose. Satisfies 1, 2, 3, 4.
- Mavis Staples (& Jeff Tweedy) - 2010's You Are Not Alone (and others). 1, 3, 4.
- Wanda Jackson (& Jack White) - on 2011's The Party Ain't Over. Satisfies 1, 2, 3, 4, 5 & perhaps soon 6.
- Shirley Bassey (& the world) - on 2008's The Performance. Satisfies 1,2,3,4,5, and of course 6.
- Neil Diamond (& Rick Rubin)- on 2005's 12 Songs. Satisfies 1,3,4,5 & 6.
- Vashti Bunyan (& Max Richter & Animal Collective!) - on 2005's Lookaf-tering. 1, 2, 3, 4, 6.

- Bettye Lavette (& Joe Henry) - on 2005's I've Got My Own Hell To Raise. 1,2,3,4,5.
- Willie Nelson (& Daniel Lanois) - on 1998's Teatro. Satisfies 1,4,6.
- Willie Nelson (& Ryan Adams) - on 2004's Songbird. Satisfies 1,2,4,5,6.
- Howlin Wolf (& Norman Dayron) - on 1971's The London Sessions. Satisfies 1,2,3,4.
- Muddy Waters (& Johnny Winter) - on 1977's Hard Again. 1,2,3.

Embedding forbidden, but click here. The best single blues session?

- Leonard Cohen (& Sharon Robinson) - on 2001's Ten New Songs. 2,3,6.
- RL Burnside (& Jon Spencer) - on 1996's A Ass Pocket of Whiskey and others. 1,3,4.
- John Fahey (& Jim O'Rourke) - on 1997's Womblife. 1,3,4.
- Glen Campbell - 2008's Meet Glen Campbell, cover dreck. 1,3,5.
- Gil Scott-Heron (& Richard Russell) - on 2010's I'm New Here. 1,2,3,4.
- Roky Erikson (& Will Sheff) - on 2010's True Love Will Cast Out All Evil. Satisfies 1 and 3.
- Candi Staton (& Mark Nevers) - 2006's His Hands. 1,3,4,5.
- Robert Plant (& T-Bone Burnett) - on 2006's Raising Sand. 1,2,3,4,6.
- Jimmy Cliff (with Tim Armstrong!) on Rebirth (2012) 1, 2, 3, 4
- The Stooges (& Steve Albini) on The Weirdness. 1,3,6. [No.]

Tom Jones continues to try, but he didn't rise anywhere in the first place, and so did not fall, and so cannot be renewed.

```

<h3>Alt-washing</h3>
<div>
  A less exalted mirror image of these albums, though: pop stars having one album produced
  <br><br>
  Tim Armstrong (P!nk)
  <br><br>
  Howe Gelb (KT Tunstall)
  <br><br>
  Bill Laswell (Motorhead, <a href="https://www.wikiwand.com/en/Brain_Drain_(album)">Ramon
  Odd choice, I grant you: but the point is that, however revered they are, these are two
  <br><br>
  Jon Brion (Sky Ferreira)
  <br><br>
  Steve Albini (The Cribs)
  <br><br>
  Glyn Johns (Linda Ronstadt)

```

<br><br>  
I could've made this list easier by just doing "surprising producers" - John Darnielle a  
</div>

# Economics as philosophy of life

Gavin

2011-10-23

```
{% include econ-life/links.md %}

<br /><br />
<!-- -->
<!-- -->
<blockquote>
    ...nobody can be a great economist who is only an economist - and I am even tempted to a
</blockquote>
- FA Hayek
<br><br />
<blockquote>
    The first question... is this question of how far life is rational, how far its problems
</blockquote>
- Frank Knight
```

Despite appearances, there is humanity in economics. Here I try to take maxims from theorems and wisdom from narrowness: together they make for a broad, honest, and inspiring worldview, nowhere near as sterile as what the field is thought to instil (as sterile as what it instils in the average student).

It's not that economics constitutes a complete worldview. But the sterility and absurdity we see in it is the result of overreach and parochialism in a few proponents, and not anything about the subject matter or even the method.

## 1. It is hard to change people.

People change all the time, but trying to direct that change is notoriously technical and intensive work. This is why some people say, mistakenly, that incentives are the core of economics: they're just the easiest way to get folk to shift. (As always, McCloskey gives a poetic rereading of an apparently boring thing: "All that moves us without violence, then, is persuasion, the realm of rhetoric.")

Take the environmental policy brouhaha - even when reasonable doubt is ruled out, when the hypothesis has attained consensus in the educated world - we keep dumping. Appeals to reason have convinced very few of us to make significant

changes. Hence, most of the large structural proposals involve increasing emission costs one way or other, and then letting people reallocate around that. Whether this is because we're hardwired for myopic behaviour by biology or psychology or culture is besides the point.

Note that this maxim does not preclude the attempt to engineer society (i.e. progressive politics). But along with #2, 3 & 7, it reminds us of the trial and error it takes.

Giant thesis: Non-political factors are more powerful than political factors in the determination of the state of the world. (But economics is only one of the non-political factors.)

Many economists give in to "It is hard to change people". The remainder of us risk making what Adrian Leftwich calls the "technicist fallacy": the dubious assumption that all governance problems have a policy solution.

## **2. It always depends.**

Economies are 'complex' in a hard sense: economic analysis takes place under gross uncertainty and necessarily limited experimentation. So unconditional answers are dishonest; it always depends. (This is not a weakness: Physical law also depends.)

Now, the third thing you learn in basic economics is the phrase *In ceteris paribus*, the assumption that only what you're looking at varies, or matters – i.e. "it doesn't depend!". But that means they admit there's a problem: it's at least explicit ignoring.

We rarely have enough scepticism. And economics is among the more sceptical disciplines: sceptical about social reality, cheap talk, professed preferences, about actual adherence to ideologies when they cost us things. Outward scepticism, anyway: as usual it's not evenly applied - you're much more likely to see radical scepticism about moral or collective action than scepticism about market allocations or the policy relevance of basic linear models.

## **3. Things fall apart; sometimes they fall into place.**

The ghost of Kant gums up arguments on political economy: many of us have the vague intuition that the amoral intentions of markets trump any accidental good that comes of them. You hear things like "capitalists don't care about social outcomes – all social outcomes determined by capitalists will be to their advantage". Well, yes, if they're doing their job and are lucky, it will. Less unreasonable is whether it is only to their advantage. This mindset holds exploitation to be any case in which people are used as a means.

(Stronger definition: "the act of using labour without offering adequate compensation". Broader definition: "any relationship of unequal benefit".)

Under these definitions, every employer is an exploiter, since they wouldn't employ you if they couldn't milk more value out.

<blockquote>

The only thing worse than being exploited by capitalism is not being exploited by capitalism

</blockquote>

– Joan Robinson

But this can't be inherently or even generally wrong: there *can be* capability and existential relief in job creation, regardless of what the employer intended. Sure, let us refuse to use people - except that my participation in this economy and that history made that move for me. My conception of what is moral has to be larger (sadly aposteriori as well as tritely virtuous).

Consider this (if it makes you angry then the ghost of Kant is in you): “the dastardly and amoral oil cartel OPEC have done more to slow global warming than all activist efforts combined.” (The argument is that by distorting the oil price upwards for forty years, they made people economise, and so incentivised the development of cleaner energy. Shoddy discussion here.) Entirely accidentally - a thing fallen in place.

### (2+3). Protection is sometimes unsafe.

The unpredictability of large-scale human affairs and the occasional emergence of order without giving orders mean even left-wing economists have to worry about our policies. Moral judgments tend to be one-step:

- “People are poor? Oh. Give em money.”
- “People pollute? Oh. Make em stop.”
- “Landlords charge too much? Make em stop.”

But the world is anything but one-step! The analysis of behaviour in terms of incentives - for all that it often justifies self-congratulatory cynicism - is at least capable of looking ahead, a little way beyond the first domino. Actually moral action demands it.

## 4. People aren't stupid

By this I mean the assumption of economic rationality. This “rationality” is quite different from the real thing, note - it corresponds to the will to more stuff and the rarer, derived will to efficiency. The assumption is a ridiculous caricature of human inner life. There's two ways for theory to succeed: either it's true, or it'd be good if it was. Since rational choice is neither, it is rejected and despised.

The kicker comes when we consider the alternative assumption: that “people are often irrational”. How do we shape policy around this? What kind of road do we build? How do we design insurance schemes or benefits? It turns out that

it is punishingly hard to do without: #4 is the behavioural principle of charity. Rational choice “theory”, reconstructed this way, is not a substantive theory at all, but a dummy methodological principle.

Now, the behavioural economists will inherit the earth soon. But policy prescription won’t easily follow from their discoveries regarding our many perversities - because while there’s ~only one way to be economically rational, there are uncountable ways to be irrational.

How can rational choice accommodate macro events like the 2008 financial disaster? Surely that really was the lord of the flies set loose in stock exchanges? In part, yes. But the good choicist’s answer is to decouple rationality from efficiency; it is in the deluded conflation of the two that the malfeasance lies. If there is no necessary link between the two, crises can be explained in terms of rational but revoltingly inefficient collective action problems, rather than by positing mass hysteria or stupidity and so getting sad.

<blockquote>

Never ascribe to malice that which can be explained by stupidity. Never ascribe to stupi

</blockquote>

&#8213; Robert Heinlein & Buck Shlegeris

The egalitarian conservatism that can be read into “People aren’t stupid” also explains why few economists take false consciousness seriously. The processes that generate our “metapreferences”, like social conditioning, are ignored. The upside of this is that economists are able to respect people’s choices in a flawed world. This is a kind of courtesy: “You’re prudent until proven otherwise”. Unlike Marxism and the new economics of happiness, even the nastiest neoclassical theory does not presume that it knows better than you what is good for you.

Ideology is too powerful and illiberal to ignore; sometimes people really are in the grip of terrible ideas / norms. You just have to recognise that there’s a cost (and a large epistemic risk) involved in calling people, or stupid people, or most people brainwashed.

## 5. You are the system.

<blockquote>

Economists are often accused of believing that everything – health, happiness, life its

</blockquote>

- David Friedman

The commodity view of existence is disturbing. Economists have viewed healthy life as a stock of capital to offer for sale (aka “labour”); babies as the investment capital of the poor; immigrants as human pollution; and any outcome below the utter numerical maximum that you squeeze out as a loss (“opportunity cost”).

There’s obvious reason to think that this framework does harm when it becomes commonsensical. But provided it’s kept contained as one perspective among many,

the commodity perspective has some important moral and policy implications: Every pound you spend is a vote for whatever you're buying. Every seven pounds you spend is another hour of your life sold.

## 6. Efficiency is humane.

Somewhere along the way in rejecting Victorian bullshit, an idea arose that being efficient is inimical to humanity. (The human will to piss about, perhaps.) This is agreeably romantic. But, in losing its social prominence, efficiency lost its moral connotation as well. (The word "economy" originally meant good household management, "thrift" comes from the same root as "thrive".)

This loss of moral charge is a mistake: the economical is ecological! Simple waste and planned obsolescence account for huge amounts of the pollution and price hikes in the world. If you ain't using it, someone will; if you don't need it or particularly want it, don't use it. And more: in high-powered contexts, efficiency saves lives, and the rejection of efficiency in the name of sweet warm human imperfection is, here, inhumane.

## 7. Sometimes there is no right answer.

<blockquote>

The curious task of economics is to demonstrate to men how little they really know about

</blockquote>

- Hayek

A common idea: "capitalism sucks but it probably sucks less than the other current options".

Since we are talking about the replacement of capitalism on capitalist keyboards paid for with capitalist pounds: capitalism obviously doesn't totally stifle future systems.

And remember #3: it accidentally clothes and feeds us, it accidentally enables state spending on education and health and law. It was forced to grant us surplus time in which to think, sometimes in which to think about alternatives. For all else that it callously does, do not deny this.

## 8. Most things fail.

Even before we consider De Beauvoir's more fatal sense: things don't work. Worse, most fail silently, creating a false sense of security. Watch its space.

<blockquote>

Did you ever think that making a speech on economics is a lot like pissin' down your leg

</blockquote>

- Lyndon B Johnson, <a href="{{jkg}}>supposedly</a>

<!-- -->

# On calling people brainwashed

Gavin

2012-03-09

{% assign adorno = “<https://aeon.co/essays/against-guilty-pleasures-adorno-on-the-crimes-of-pop-culture>” %} {% assign orlando = “<https://books.google.bs/books?id=zfuwDwAAQBAJ&pg=PW4TURYYD&sig=ACfU3U1Kn1cbn7TZpSTt5zhGv43lG6K9PA&hl=en&sa=X&ved=2ahUKEwiDi5OBmv1AhVSTDABhO2An4Q6AF6BAGFEAM#v=onepage&q=%22stultifying%20effects%20of%20popular%20cult%22>” %}

I can only suggest that he would combat false consciousness to awaken people to their true interests has much to do, because the sleep is very deep. And I do not intend here to provide a lullaby but merely to sneak in and watch the people snore.

– Erving Goffman

Never ascribe to laziness that which can be explained by people knowing their own lives better than you do.

– Buck Shlegeris

Freedom is not simply the right of intellectuals to circulate their merchandise. It is, above all, the right of ordinary people to find elbow room for themselves and a refuge from the rampaging presumptions of their “bettters.”

– Thomas Sowell

## First-order psychology

- electoral interference.
- Body image,
- Violent videogames
- Porn
- Advertising

What is this idea that the Russians can spend \$0.1m and totally pervert American discourse, but that Michael Bloomberg can spend \$700m and die on his feet? (Do the Russian really understand America 10,000 times better than the Bloomberg campaign? Does taking the gloves off allow you to do this much better?)

Hegemony. Chomskyan consent.

Foucauldian biopower.

## **Radicalism implies that everyone is brainwashed**

Many interesting theories accuse us of being brainwashed: “You lack information; now, open your eyes”:

People have this naive equivalence between a fantasy and a preference, a voluntary simulation and an increased inclination to *do*

the unconscious mind (you’re so unconscious you don’t realise you’ve a massive unconscious)

Radical feminism (you’re so oppressed you don’t know you’re oppressed, and/or you’re so sexist you don’t know you’re sexist)

Or more generally Critical theory (you’re so oppressed you don’t know you’re oppressed).

eliminative materialism (you’re so evolved you don’t know you’re evolved)

Wittgenstein II (you’re so linguistic you’re constantly lost in the woods of your words)

genealogies of anything (you’re such a slave you don’t know you’re a slave),

Whether the culprit is folk theory, oppressive social structures, blind academic paradigms, or just our own narcissism, we’re told that our intuitions make fools of us. that our attitude doesn’t reflect our objective position. All-too-human. The conceptual brainwashing involved in all this goes by a number of names depending on the scope of the alleged ‘wash: bad faith and doublethink, false consciousness, latent ideology, cultural hegemony, Repression, ‘simple’ paradigmatic underdetermination, and so on.

Many of us love to be accused: the above theories are the biggest doctrines in the intellectual culture of the humanities. What is it about certain concept changes that make us such enthusiastic masochists? Can we only be made to listen to theory when there is a slap in it for us?

Now, in the above I conflated scientific revolutions with political identity-work. I suppose I should stop doing that and become serious, since, unlike (most) natural science, the political kind challenges us in a desperate existential way - it accuses us of misunderstanding ourselves, not just continents, disease, or motion. We’re still so conceited about our self-knowledge that we can’t help but be stung by the suggestion of programming, and stung, I reckon, to morbid curiosity and codependency.

(The difference between a belief in hard determinism and the belief that everyone is brainwashed is a subtle one.)

The effort to deprogram us can take a few targets, some more tractable than others: there's your shite assumptions (particularly prevalent in the folk theory of gender); your ignorance of the structures that you are formed and active within, whether these are causal structures (as when we imagine our magic free will in the face of neuroscience) or social (as in the popular theory of late capitalism's spooky mind-control).

- Your doublethink and your emotional habits. (e.g. knowing that you're a healthy weight but being nagged by body image issues regardless.)

So what is it that makes us enthusiastic masochists over these ideas?

Well, there's the delicious drama of it, their portentous invitation to reason; or the backup it gives to the enduringly moronic Great Man theory of history ("Thanks to Him, we now Know!") but I think the key driver is brainwash theory's indirect invitation to narcissism. For the 'masochism' of accepting really new theories is just a stage, after which we get to claim to have transcended our brainwashing, and to feel that we've joined a vanguard; a little pocket of knowledge in a corrupt and stupid world. (And: "Now to impose our will on the deluded.")

This new brainwashing - the arrogance of the self-conscious theoretical élite - is far harder to rinse away. woe betide us.

## **Everyone's at it**

It's not just the French and the Arts boys who rely on this.

The anti-postmodernist "grievance studies" / *Fashionable Nonsense* / "victim studies" view requires exactly the same kind of accusation of false consciousness as does Marxism and critical theory do:

"All these thousands of intelligent people must have fallen into an ideology that blinds them - or must just be innumerate - or must just be in a purity spiral..."

Now, false consciousness clearly exists - witness the cult of the British monarchy among basically all working-class English people But it is a nuclear option, an irreversible form of ad hominem

better to assume different values, different priors, and different aims in general And leave our critique on the level of particular bad scholars (better! Particular bad *papers*)

I foolishly wrote off anthropology and sociology in 2013-2017, and was lucky to find some remarkable, realist scholars in each since.

## How I stopped valuing radicalism itself

I remain radical in some ways. I think our ignoring the suffering of others - the global poor, nonhumans inside and outside our industrial farms, or future generations - is monstrous. I think that we should work very hard to overcome many bad parts of our biology (with longevity, gene therapy, new moral intuitions). I think that one of the most pressing problems in the world is making sure that future AI systems go well. My government has recently been involved in a number of terrible crimes. There is so much wrong with the world. A good place would be radically different, and the longer it takes us to get there, the more misery will be recorded.

But I no longer use “radical” as a compliment and no longer value the label. Thatcher was a radical; Stalin was a radical; Kaczynski is a radical. Just being strongly opposed to the status quo, or just taking direct action to change it, is not laudatory.

Two problems with treating it as an end:

- 1) it's supremely difficult for a radical to not hold most people in contempt. Largely undeserved. Election go the other way? “It's not that people have honest differences in values, ideas about the good, or about what causes the good. Instead they're rubes, or sadists, or egotists.”
- 2) you've allowed something other than either truth or goodness into your objective. Mere novelty, mere contrariness, mere extremity.

```
<h3>Been in the Wars</h3>
<div>
    Science Wars
    Culture War
</div>
<!--
-->
<h3>The pomo critique of pop and globalised pop</h3>
<div>
    The above idiotic model of psychology pops up again in the elitist postmodernist disdain
    <blockquote>
        people never passively absorb cultural messages... "The incandescence is not simply a su
</div>
<!--
-->
<h3>Epigram</h3>
<div>
    <blockquote>
        It's a curious thing, that the mental life seems to flourish with its roots in spite
    </blockquote>
    <center> - DH Lawrence </center>
```

</div>

### See also

- Mako Shen on dream arguments, false consciousness, and the nervous energy currently called woke.

# to desperately instrumentalise myself

Gavin

2012-05-29



... there is an internal ethical urge that demands that each of us serve justice as much as he or she can. But beyond the immediate attention that he rightly pays hungry mouths, child soldiers, or raped civilians, there are more complex and more widespread problems: serious problems of governance, of infrastructure, of democracy, and of law and order. These problems are neither simple in themselves nor are they reducible to slogans. Such problems are both intricate and intensely local...

- Teju Cole

Specialisation is for insects.

- Heinlein's Lazarus Long

Turns out that a degree - even one on 'real world' topics like, supposedly, economics - isn't a skill. Isn't really much to do with much. This is galling, because I have bottled action in me and have failed to get moral hydraulics to steer it.

Is that too reductive? I might not have such a quantity of good intentions without my years among the humanities; they only suck for obtaining hard skills. And 'hydraulics' means just narrow technical skills. To have those is to be able to instrumentalise oneself: to have the option of production. (More often, you're made to get credentials that imply you are productive.)

What spiritual costs does this instrumentalisation levy? I was at a conference the other day where people were banging on in the Frankfurt way about 'instrumentalisation'. I do sympathise with their background theory - which attributes modern atrocity and mental illness to the reign of scientism and the cult of practicality - but not in the uncritical, almost superstitious, way it gets invoked. Useful things are abhorrent to a certain mindset. Since they following Horkheimer who followed Kant, what I've read of Cultural Studies tends to bear an awful, watery stance, where an agent or project's being problematic implies that it's taboo, irredeemable, a moral medusa.

In discussing the 'white saviour complex', one speaker implied that objectifying someone you are trying to help is such an evil process that it negates any good

your action might cause. Teju Cole:

From the colonial project to Out of Africa to The Constant Gardener and Kony 2012, Africa has provided a space onto which white egos can conveniently be projected ... The banality of evil transmutes into the banality of sentimentality. 'The world is nothing but a problem to be solved by enthusiasm'.

This conflict leads to condemning the attempts of all kinds of liberal structures (welfare state, NGOs, the UN), and from there, passivity. Because they rightly probe the mixed motives and identify unconscious power structures in do-gooders, the scholar can feel satisfied in holy inaction. This is the accidental turn of the 'New' Left ; reading is not only political, but political enough. The only labour you owe to the disadvantaged is your intellectual labour; since everything else you might try is tainted.

But as long as it is chosen, as long as it's not the only thing you get to be, there's little wrong with objectifying yourself, choosing to become, among other things, an instrument. The trick is to retain your radical goals even with a prosaic, professional, instrumentalised exterior.

(Case in point: East Africa is chronically, catastrophically short of Quantity Surveyors. Apparently.)

---

Long story short; let's go make ourselves useful:

Knots (1 week; £minimal)

First aid (1 month; £minimal)

Driving (4 months; £400)

Databasing. (a month or so; £2000)

PGDE (1 year)

MA African Studies in Nairobi or Makerere (1 year; £1000)

MSc Maths, Open University (takes 2 years part-time; £2500)

MSc Dietetics, QMU (2 years pt; £4000)

SVQ Mechanicking (just motorbikes, probably; 2 years pt; £1000)

ACA Chartered Accountant (for NGOs, taking the ICAEW qualification, 2 years pt)

Chinese (3 years in-country - cf. TEFL; -£2000)

2025, maybe:

PhD in Irrationality (designing cognitive bias education programmes)

or Development (new metrics and meta-analysis for aid dependency)

or Animal rights law  
or Nutrition/Biochemistry (on the prospects of nootropics)  
or Transhumanism in general (on theodicy and the love of suffering)  
or Epistemology (radical scepticism's influence on contemporary philosophy)  
or Poetry (contemporary developments, or lack thereof)  
or Metaethics (on problems with Humean sentimentalism)  
or Nationalism (the idea of a national 'mentality' esp. Scottish)  
or Economic methodology (statistical/empirical tests of the most sophisticated models of fiscal impact)  
or Econophysics, University of Houston

# Existential overheads

Gavin

2016-08-24

```
{% assign traz = "https://twitter.com/MichaelTrazzi/status/1210242894323503104"
%} {% assign sleep = "https://www.philips.com/c-dam/b2c/master/experience/smartsleep/world-sleep-day/2020/2020-world-sleep-day-report.pdf" %} {% assign comm =
"http://www.worldmapper.org/posters/worldmapper_map141_ver5.pdf" %} {% assign shower =
"https://www.bbc.com/news/science-environment-15836433" %} {% assign shop =
"https://qz.com/1677747/americans-are-spending-way-less-time-shopping/" %} {% assign g =
"https://www.gwern.net/newsletter/2019/13" %} {% assign work =
"https://en.wikipedia.org/wiki/Working_time#/media/File:Heures_travailles_OCDE.png" %} {% assign ass =
"https://web.archive.org/web/20150314181445/http://www.theweinerworks.com/?p=1694" %} {% assign comp =
"https://eatcomplete.co" %} {% assign cook =
"http://www.statista.com/statistics/420719/time-spent-cooking-per-week-among-consumers-by-country" %} {% assign exercise =
"https://www.nhs.uk/live-well/exercise/" %}
```

Average time used for basic upkeep of the organism 2, on a weekday:

7 hours sleep. (29%)

7 hours production (29%)

1.5 hours commute (8%).

1.5 hours cooking/eating. (8%)

0.5 hour hygiene (2%)

0.3 hour exercise (1%)

0.3 hour shopping (1%)

= Leaving 6 hours for actual, discretionary life (25%).

The above is not exactly waste, since each of them have their own pleasures, since some fraction of people would perform their jobs even without pay, and since (unfocussed, mostly non-meaningful) cognition continues throughout them.  
3 But it is still unfree.

Three possible reactions to the realisation that 75% of your time is not wholly yours:

- *Quiescence*. Many people seem to spend their 25% on screens and tidying.
- *Mindfulness*. Maybe what's bad about the above ratio is in our head, and maybe close attention to the world around us can make the above meaningful.
- *Rage*. Fuck that: Optimise, race, and cut.

## Against the dying of the day

expecting a large increase in the average (treating it as a latent variable in being spread across many measured variables) is entirely missing the value of productivity. It's not that one gets a lot done across every variable, but one gets done the important things. A day in which one tidies up, sends a lot of emails, goes to the gym, walks the dog, may be worse than useless, while an extraordinary day might entail 12 hours programming without so much as changing out of one's pyjamas.

- Gwern Branwen

How do we get life back?:

1 hour off commute by taking public transport (used for reading)

(Or + 1.5 hours off commute by working from home)

(Or + 0.5 hour from cycling your commute)

Maybe 0.5 hour saved on sleep from oral melatonin.

0.5 hours saved on lunch from having a ‘complete meal’ shake.

0.4 hours off shopping from home delivery

0.2 hours by taking caffeine & theanine in pills instead of boiling decoctions.

7 hours by finding work you really think should be done. And by you.

10 hours from becoming a crusty freegan (8 hours off work 1.5 off commute, 0.5 off hygiene).

1

The most effective strategy for preventing waste of life is hard to quantify in terms of hours per day: it is the behaviour implied by the expression “proceed til apprehended”. Job requirements are often nonsense. Surveillance is (so far) gappy. Guards are largely indifferent. Meetings can usually be skipped. Some red tape is purely decorative: not even the demanding authority thinks it matters.

4

```
<li class="footnote" id="fn:1">
<a href="http://xkcd.com/1205/"><i>Objectively, should I care?</i></a>
```

```
</li>

<li class="footnote" id="fn:2">
    This is a childless developed-world person, clearly. In large parts of the world was
</li>

<li class="footnote" id="fn:3">
    Most delightfully <a href="http://showerthoughtsofficial.tumblr.com/">shower thoughts</a>
</li>

<li class="footnote" id="fn:4">
    There are situations where the spell is inapplicable, like anything to do with the police
</li>
```

# Virtue, work, and the world to come

Gavin

2014-01-10

```

<!-- -->
<br>
<span style="font-size: small;">
    (c) <a href="{{how}}>Meghan Howland (2012)</a>, "<a href="http://www.greynotgrey.com/bi
</span>
<br>    <br>
<!-- -->
<blockquote>
    <i>A rich boy goes to college. He makes a lot of friends. They all think they are specia
</blockquote>
<div style="line-height: 5px;">
- Sam Lipsyte<br></div>
<br><br>
<!-- -->
<blockquote>
    <i>Freedom in an unfree world is merely licence to exploit</i>. <br>
</blockquote>
<!-- -->
<div style="line-height: 5px;">
- Germaine Greer
</div>

<h3>Disclaimer</h3>
<div>
    <i>May 2020:</i> This is a bit of a mess. It makes no sense without the background, that
    I'm posting it because it represents a huge philosophical shift, one which defined the m
</div>
```

Reportedly, the ‘only really serious’ philosophical question is whether or not to kill yourself. If we take the point of this to be that your answer to the suicide question might preclude you answering any other questions, and if the importance of a question is somehow transitive with the importance of questions it affects, then that’s sort of true if you squint. But it is much more likely that you’ll currently be faced with slightly less stark choices. Like: What will you (try

to) do? Where will you do it? With whom? “What will you do?” is the tough one: the other two usually follow in a straight line from it, even now, unless you choose to do nothing, or choose something that’s wanted everywhere, like medicine, or web development, or having a really good flow. But surprisingly few people explicitly think about these: instead you just fall into the job that happens to be going, and then you stay there if you can. (Or for academics: you slip into a degree based on your highest grades at school, and take on that field’s fixations, and stay there if you can.) You live where you’ve always lived; and you go out with who you can. A default decision tree of life might run like this:

- 1) What will you do with your life? Don’t really know. Money. House. Couple kids. Hobbies.
- 2) Where? Eh? Here. It’s where all my stuff is.
- 3) With who? Well, with my mates, and Mr/Ms/Mx Right.

To get some help with (1), I spoke to the blue-sky pragmatists at 80,000 Hours about making myself useful (taking the ‘effective altruism’ omnibus). They’re an ethical-career research group offering what you might call the Engineer’s Guide to Moral Transcendence: they work by appeal to economics, cognitive science, and an arch-consequentialism. For people with the stomach for it, they recommend indirect altruism - things like ‘Earning to Give’, getting yourself a high income so as to sustain a high volume of charitable donation - as a surer, magnified way to benefit the world. This is because when we adjust for psychological availability, counterfactuals, and prestige, the effects of actions often turn out counterintuitive.

80k have a single pledge of membership: “I intend, at least in part, to use my career in an effective way to make the world a better place.” Inevitable value conflicts aside (“better according to who?”), this is as good as anything so general can be. So:

- 0) Will you live? Oh go on then.
  - 1) What will you do?: Help out.
    - a) How? Professional effective altruism; doing as much good as possible.
    - b) Which ‘ethical’ career exactly? Hm. Give me a minute; let me run the math.

Previously, I was inclined to just get myself two Professions (one pro bono professional role, and a moonlight public intellectual role), research a ruined geographical area, and get stuck in. (The initial list was public statistician / bioethicist / Teacher / Accountant / Dietitian.) There is a shortage of meaningful jobs. This is probably not because people don’t want to do them (the average British third-sector vacancy receives x application), but because there isn’t the funding. Thus there is actually something *prima facie* wrong with ploughing on with that UN or whatever.

Three good reasons to Earn to Give, then:

- 1) It makes an actual difference. Terrible correlate!  
My labour is replaceable; in general, it just crowds out other people’s.

My donations are not.

- 2) You can fund multiple workers.
- 3) Not just preaching to the choir; into lucrative industries who are more likely indifferent and full of disposable income.

[http://www.academia.edu/1557895/Against\\_the\\_Common-Sense\\_View\\_of\\_Ethical\\_Careers](http://www.academia.edu/1557895/Against_the_Common-Sense_View_of_Ethical_Careers)

<http://oxfordleftreview.files.wordpress.com/2012/07/issue-7.pdf>

[http://lesswrong.com/lw/cxj/debate\\_between\\_80000\\_hours\\_and\\_a\\_socialist/](http://lesswrong.com/lw/cxj/debate_between_80000_hours_and_a_socialist/)

[http://lesswrong.com/r/discussion/lw/fkz/responses\\_to\\_questions\\_on\\_donating\\_to\\_80k\\_gwwc/](http://lesswrong.com/r/discussion/lw/fkz/responses_to_questions_on_donating_to_80k_gwwc/)

The unsexy and philosophically suspect mechanisms, maximisation and prioritisation take on enormous significance when lives are at stake.

---

'Making a difference', if it means anything, means bringing about good things that wouldn't have occurred otherwise. But then when people think about which careers are ethical, they often seem to focus on which careers do good directly – doctors, aid workers, campaigners etc... we want to bring about positive consequences that wouldn't have happened otherwise: to really make a difference.

– Benjamin Todd

By connecting this vision with the mechanisms of labour and capital, 80k also raises a more systematic problem. In the face of grave ethical demands, how are our choices structured? Do we face a world where choosing a career is the most important decision we make? Or does this individual dilemma obscure a more complex and perhaps more contingent reality?

– Tom Cutterham

This hyperactive liberal humanism faces a lot of vitriol from the Left, though. Why not be a philanthropic banker? I've written an FAQ for this to prevent: none of the objections are fatal and many seem to me to be hollow applause lights. Longer pieces criticising EA:

Insider.

Debate, one side ignoring probabilities.

\*\*\*\*\*

Effective Altruism is part subversive, part conformist: subversive in its radical egalitarianism and its critique of complacent privilege; conformist in that it's another force channeling us towards the traditional success model... the ironlogic of replaceability leaves many dreams dead on the ground, to be sure. But is this a problem with EA as an ideology, or a problem with reality?

– Rhys Southan

Further to not working in banking, but for non-ethical / non-political reasons:

It's boring.

The attitudes of the typical finance colleague are disgusting,  
the common language of the whole professional sphere is disgusting,  
the actions it enables are disgusting.

Worst of all, there's usually a vast split between the things that make you a good person and the things that make you a good worker. What we do at work leaks into our real lives. I'm determinist enough yet to admit that my surroundings can and will morph me - & who wants to be morphed, in habits, inner life and reference pool, into the compleat Accountant?

Sorted: no then. The gist of this debate about effective altruism is: "you can't do good without political engagement too". But social movements are also problematic. Further to not being a Trot:

We are working at the margin; we always base our decisions on the state of the rest of the structure and the tractability of the problems.

We maximise because, if one is undertaking a really effective type of action, small extra improvements can make a difference to hundreds of more lives.

Group dynamics (such as are found in social movements) distort us deeply: politics is the mind killer.

Group dynamics are also really boring.

Dhaliwal is obviously spot on about me and my sort: we have asked "what I can do?", rather than "what can we do?" I take this to be an argument for returning to my original plan, of a direct career in something helpful; while I eschew the larger and more tasteless transformative political work, if I can give my life to direct work for the oppressed, I'm still not detached and contemptible. Is that really solidarity? Not sure. Am I on the wrong side of history, then? Hardly. In the same way that no god who would punish me for my warranted doubts is worth abasing myself to anyway, no revolution who'd end me for deciding against their quixotism is worth fomenting. I want to be neither a high-impact shill nor an endless-vacation revolutionary. Thus do I learn the limits to my altruism: namely boredom, and suits (clowns to the left of me, jokers to the right).

ANALYTIC EGALITARIANISM Things we have tried:

Give 'em a bunch of stuff.

Send some people to see what needs doing (Technical support)

Give their governments a bunch of money.

Pros: Not remotely imperialistic. Prima facie efficient.

Cons: Corruption. Fungible with arms spending to an enormous, lethal degree (11%).

Ask 'em what they want and give 'em a bunch of that stuff.

Empirical development.

Among other things (mostly health interventions), turns out that “give ‘em a bunch of money” is a solid move.

---

They sentenced me to twenty years of boredom for trying to change the system from within. I’m coming now, I’m coming to reward them.

– Leonard Cohen

The simplest objection to large giving pledges is simply that your money is yours: “I sold my labour to obtain this; so I get to decide what to do with it, no-one else has a claim to it.” Textbook economics backs this territorial claim: “wages are just, because we are paid according to our relative productivity.” But the little-known fact of the matter, though, is that very little of your total wage is determined by your particular skills and negotiation. 3/4 of the average developed-world wage is a direct result of the wage level of the society you happened to be born to. In a strong sense: you are overpaid.

Most people in rich countries get the wages they do now only because they exclusively share their labour market with some very productive people, who outperform their counterparts in the developing countries by hundreds, or even thousands, of times... this kind of wage gap cannot be justified – how can you have two people doing the same job with equal efficiency being paid wages that are 20, 50, or even 100 times different.

– Ha-Joon Chang

This kind of moral luck is not tolerable. What, then, is the highest wage one can justifiably allocate oneself? Or, to put it in a less loaded way, what is ‘optimal’? I have an argument about this here.

\*\*\*\*\*

With the caveat that I’d write critically about them, I joined Giving What We Can, anyway. With the caveat that it’s messier than it looks, so should you.

# Can you trust your methods?

Gavin

2015-05-08

```
{% assign jaynes = "https://en.wikipedia.org/wiki/Mind_projection_fallacy"
%}
```

If a person knows he is being denied an opportunity... he can never be quite certain whether his lack of desire for it is shaped by the fact that it is unavailable to him ("sour grapes"). That gnawing uncertainty counts as a harm.

– Jon Elster

ONE hot summer's day a Fox was strolling through an orchard till he came to a bunch of Grapes just ripening on a vine over a lofty branch. "Just the things to quench my thirst," quoth he. Drawing back a few paces, he took a run and a jump, and just missed the bunch. Turning round again with a One, Two, Three, he jumped up, but with no greater success. Again and again he tried after the tempting morsel, but at last had to give it up, and walked away with his nose in the air, saying: "I am sure they are sour."

– Aesop

A research programme that would be very illuminating and very unpopular: How much is someone's methodology to do with rationalising their particular abilities?

Does not having the skill to conduct either quantitative or qualitative research correlate with denying its value? (Clearly yes.)

Given that people very often adjust their desires to their opportunities, and given that methodology should ride on higher things, I propose a trio of studies to check the academic community's hygiene:

- Sour symbols: Disparaging or emphasising the limits of quantitative reason because you yourself are bad at maths.
- Sour mouth: Disparaging or emphasising the limits of qualitative reason because you yourself are bad at criticism or phenomenology.
- Scoundrel bastions: What fields do people with *neither* competence flock to?

The inverted forms – seeing what you’re good at as a superior insight into the world (“sweet lemons” and “mind projection”) – are as important, but hopefully get captured in the first correlation.

One could use the SAT or GRE to obtain a proxy of verbal and mathematical reasoning ability; people would object to this, 1) rightly because timed tests are an artificial measure of research ability – they can prove ability, but they can’t really disprove real-life ability – and 2) wrongly, because it threatened their status.

By combining the two studies within-subjects, we could derive a general factor of adaptive methodology: how much a given person is swayed by their own lack of skill. This could be a proxy for how rationally they conduct themselves in general.

---

I respect Putnam’s and Rorty’s criticisms of positivism because I know they are profoundly skilled in logic; I trust Deirdre McCloskey much more in her postmodern libertarian feminism because she was both a quantitative historian and a socialist in her youth.

---

Good methodology can substitute for brilliance: if you follow the scientific method long enough, you will find stuff out, almost regardless of your acuity or creativity.

An unfortunate demonstration is Thomas Midgley’s discovery of tetraethyl lead: “At war’s end he resumed his search for a gasoline additive, systematically working his way through promising elements in the periodic table, and in 1921 he and his team found that minute amounts of tetraethyl lead completely eliminated engine knock.” Four years of dumb permutation!

If you can understand an algorithm’s steps, you can perform incredibly complex mathematics given only patience and a pen. (Or wings.) In programming, object-oriented languages enforce a simple stepped method that allows numpties to make, well, most of the internet.

Relatedly: to have the studies produce results of lasting worth – rather than results for wreaking retribution on idle methodologists – we’d want to track the things that practitioners did. (Though is there any such thing as a practitioner, in philosophy?)

---

My saying ‘methodology’ in the above makes the point seem irrelevant to anyone but academics or devoted autodidacts. (The word only really denotes the formal and contrived ways that we act when we know we’ll have to face scrutiny.) But the implications go way beyond those islands in the sun to the grody places in which most thought lives.

Computer science: the methodology is necessarily quantitative.

Philosophy: methodology largely qualitative (though with a distinct subculture of utter quants / meta-quants). Everyone's a methodologist.

# The universal love of suffering

Gavin

2015-02-11

```
<br>
<span style="font-size: small;"> 'The Marriage of Heaven and Hell', (c) Keith Haring (1984)
```

Begin by lining up our suspects:

```
{% include suffering/exemplars.html %}
```

What marvellous consensus! Taoists, Stoics, conservatives, radicals, Christians, Jews, Fascists, captains of industry, artists. United in this! Call the doctrine they hold in common contrast ideology. "Tajitism" might seem a fairer name, after the famous yin-yang symbol - but the Asian doctrine is actually not the earliest of the symbol's tendrils. (Not that identical symbols imply identical ideas.) Also, Contrastivism took that word already.

Contrast ideology is one of the most interesting ideas I know. It leads to the most fruitful theodicy, a fascinating semantics, and also to the most therapeutic secular philosophy of life. It is a deeply beautiful thought - and I will fight it until it dies. (Or until it allows alternatives to live.) I'm here focussing on the most pressing subset of contrast ideologies: the philosophical justification of suffering, call it hormetism or lacrimism. It is relativism with regards to suffering; the pessimistic belief in the symmetry and obligate symbiosis of the polar values. I there associated it with structuralism - but rather than linguistic, literary or anthropological structuralism, lacrimism is naive ethical structuralism. It needs attacking because it promotes conservatism and indifference to world problems - often on matters that aren't intractable. It also weights a portion of the world down with a piteous, gnarling masochism (which, I think, outweighs its therapeutic role). Less seriously, it underwrites a repugnant conception of art: as dark Romantic summum bonum. The world is rammed full of suffering. How can we cope with this?

Sado-masochism.

Withdrawal (asceticism; indifference)

Lacrimism

e.g. Stoicism

e.g. Theodicy

Abolitionism.

The rest of this essay suggests that all religious and philosophical responses follow one of the above.

i.

I have perhaps been muddling several claims together, one or two of which are true and good:

“The world is essentially just.” (Just-world. Karma)

“Suffering is profound”

“There are other worthy goals but happiness”. (Yes!)

“Positive and negative things are not alien to each other, but related and continuous.”

Suffering and only suffering enables defiance.

Suffering and only suffering produces strength.

Conflict and only conflict produces progress (Mark and Hayek!)

A series of well-documented scientific (hormetism; value relativity; experience-stretching). I can’t oppose.

Pains that lead to gain are comparatively rare. And that it is an inverse relationship: (“Its insistence on the evil in man’s nature, and in particular on the root of that evil, suited the New England temperament well which had been shaped by a similar Puritan emphasis. In fact, to hear Anna Freud speak of the criminal tendencies of the one and two-year-old is to be reminded inevitably of Calvinistic sermons on infant damnation.” - David McClelland.) Biological hormesis is relatively new but an extremely well-confirmed natural phenomenon. The hygiene hypothesis is the most famous instance, and mithridatism the most sensational. (Didn’t your school friends speak of Rasputin with awe?)

0. Stoicism. a. Trivial Lacrimosa Principle: Good can come of suffering; it’s not always in vain. (It can have extrinsic value.) b. Strong Lacrimosa Principle (descriptive): Negative experience is essential to the existence of positive experience. Good would not be good without bad to identify it against.

b\*. SLP (prescriptive): Suffering promotes the existence of great things. c. Ultra Lacrimosa Principle: Suffering is intrinsically valuable. Suffering cannot be in vain.

d. Nihilist’s Lacrimosa Principle: The positive and negative are so mingled that distinction between them is illusion. Happiness has no intrinsic

I am a fool if I deny the weak principle. But the lessons I derive from it are very different from lacrimism’s; and our acceptance of stronger forms must be only as a philosophical last resort. We might yet change things enough.

if, by the spirit, you put to death the deeds of the body, you will live." - Paul of Tarsus, the passage that supposedly mandated whipping yourself

ii.

```
<blockquote>
    that so much suffering can be in vain is intolerable to me, it kept me awake all night:
</blockquote>
<p style="line-height: 5px">
- Andre Gide</p><br><br>
<!-- -->
<blockquote>
<i>A good Westerner, Andre Gide couldn't help but think that suffering was the price of happiness
</blockquote>
<p style="line-height: 5px">
- Eugene Ionesco</p>
```

We owe lacrimism to theology. It came of the brilliant thousand-year intellectual project to let God off the hook of criminal negligence.

Philosophies less leashed to an assumption of inevitable cosmic justice, however, The Just-world hypothesis Pain breeds self-pity more often than empathy or strength.

iii.

```
<blockquote>
    The great pre-industrial and pre-scientific civilisations, especially perhaps the Western
</blockquote>
<p style="line-height: 5px">
- Ernest Gellner</p><br><br>
<blockquote>
    [Freud's unconscious] <i>wish to be morally censorious about humanity, a desire to make
</blockquote>
<p style="line-height: 5px">
- John Wren-Lewis</p>
```

Most commonly a submerged religiousness. "We are fallen, so we deserve to suffer."

Accusing anyone of rationalisation is always shaky. But isn't the strong lacrimosa principle precisely the story that an entity impotent in the face of lifelong angst and pain would tell itself, to get by?

iv.

Antithesis. Our prejudice for theoretical symmetry has caused a great deal of harm already. Unify.

Kierkegaard holds despair to be central to human nature. Uncertainty.

v.

My life consists in my being content to accept many things.

- Wittgenstein

So, what's the alternative? Don't I have to be radically manichaean? Am I not falling foul of what Nietzsche flagged up ages ago as the irrational love of antithetical values? Will I destroy lacrimism's placebo effect in the name of truth? Can one really reconcile this moral dualism with one's love of shiny pluralism?

Disclosure: Some time ago, I watched someone die of oesophageal cancer, over the course of about a year. The experience was not salutary to his spirituality, personality or worldview, nor to mine. I write from the force of this anecdote, but not only that.

# Present pieties

Gavin

2015-06-04

{% include piety/links.md %}

... one knows a piety from a principle because even those who oppose a piety have to pretend to honour its core point.

– Adam Gopnik

... one way you know that something is an institution is that you don't have to give reasons for it. Getting a college degree, like getting married, is what people do.

– John Emerson

What does everyone have to like? (Better: what does everyone say they like, in the West, if they're respectable?) By construction, most people don't like the things hipsters like. What do even hipsters fail to react against?

Not everything in the following list is bad, just mysteriously universal or unquestioned. I also wanted things independent of Left/Right politics, because they are both obvious and plainly not universal. So I haven't included Gopnik's example (gay marriage) despite it having all the hallmarks of a piety – e.g. finding support even among ancient enemies, having only confusing opponents we could report on, gawkily.

They're also superficial in comparison to the deadly pieties, 'things one loses one's job for opposing'. For which 'heresy' might not be excessive. Though that takes our term for 'things others will kill you for saying'.

<h3>Update (2020)</h3>

<div>

To my surprise there has been some movement. It is now *much* more common to be sce

<!-- -->

One piety I missed, which was present in 2015 but has gotten much worse, is the entire w

</div>

{% include piety/journ.md %}

{% include piety/trav.md %}

{% include piety/nat.md %}

```
{% include piety/net.md %}  
{% include piety/ed.md %}  
{% include piety/read.md %}
```

### **How does this happen?**

Well, as I've conceded, each of them has some intrinsic appeal. But that aside: herding and signalling is the boring but probably most important component. Then there's marketing, which I actually don't see as that powerful. Preference falsification is cool but requires some great force that makes everyone lie without co-ordinating the lie.

Each of the above sustain an identity in their host. And once a practice gets into your identity, it can extract a huge amount from you, without you ever thinking to complain. It may seem far-fetched that someone could identify with their consumption of journalism, but behold the Extremely Online, the politics wonks, the amateur pundits on a million radio call-ins, all over the earth.

### **See also**

- Liam Bright on the link between neoliberalism (yes, really) and identitarianism.

# magic words

Gavin

2016-08-12

What is a magic word? A word which is not just a symbol? A causal word?

Such spells exist. Just say “noise” to yourself, and it appears. Shout “police!” long enough and you will summon the demon you name. “Confusion!”, in company. “Speech-act”. “Disquotation” après-moi. “Entropy!” arguably.

(All quite aside from code, the living words eating the economy and the intellect, and so the world.)

{% include padder.html howMuch=14 %}

# on veganism

Gavin

2016-08-22

(c) Lucian Tidorescu (2013)

I don't give money to places that harm animals; the biggest part of that is not buying meat, eggs, dairy, etc.

I haven't written much about it; I dislike signalling in that way and I'm suspicious of the effects of calling-myself-things on my thoughts.

There's a certain kind of vegan - anti-modern, pious, loud - that perhaps limits the audience for animal rights. So I had better pipe up with my supposedly more rational, bioprogressive / ecomodernist form. (Someone else will have to handle the task of making it not seem weird.)

I was recently interviewed on the topic for some sociology research:

Can you tell me in your own words, what your definition of veganism is? What are your reasons for being vegan?

Veganism is usually 'not consuming animal products'. My sort follows from a more general view of what is ethical: 'don't cause harm to anything which can probably experience harm'.

Vegetarianism is actually implied by common beliefs, but few people act like they've joined the dots:

It is wrong to cause unnecessary harm.

Factory-farmed meat animals suffer.

Humans do not need to eat meat to live or thrive.

Therefore eating factory-farmed meat causes unnecessary harm.

Therefore it is wrong to eat factory-farmed meat.

My brother is a meat-eating freegan, but that fits consequentialism: it is not eating meat that causes harm, but sending economic signals that eventually cause the meat industry to cause harm. Similarly, I wear clothes bought from charity shops, including leather; since reusing clothes does not constitute economic demand, it causes no animal suffering, so it is morally neutral. Or, actually slightly positive, since it obviates the production of new clothes.

I don't mind if people say I'm not 'vegan' as a result. The label is not the point: stopping the harm is. It all boils down to harm reduction:

direct harm, since the industry inflicts pain on billions of creatures, totally unnecessarily;

macroeconomic, since meat production wastes huge amounts of water, land and energy, which deprives many humans of resources and drives up food prices;

environmental, since the carbon emissions involved could eventually cause vast suffering through climate change;

antimicrobial resistance: the industry includes antibiotics in the feed of animals, to prevent the disgusting conditions from affecting output - this systematic administration squanders a very precious resource: the effectiveness of our medicines; this process potentiates:

the zoonotic risk: most human pandemics have been novel mutations in nonhuman diseases. So, by incubating billions of animals in terrible conditions, the meat industry is thus an unparalleled opportunity for global plagues.

What does it mean to you personally?

It is a minor chore I undertake in order to meet minimal ethical standards: first, do no harm.

Can you remember a specific moment that triggered your veganism? What was this?

I'd been vegetarian since I was 16, for utilitarian and anti-capitalist reasons (e.g. McDonalds deforesting Brazil for grazing). In a first year ethics course, my lecturer pointed out that ethical vegetarianism is hypocritical:

since the dairy and egg industries are either the selfsame companies that form the meat industry, or their operations share profits and harmful processes with the meat industry;

since surprising amounts of harm are an essential part of even nonlethal factory farming (e.g. male chicks electrocuted to death at birth, calves separated from mothers at birth).

Environmental vegetarianism is also inconsistent: cow's cheese causes more CO<sub>2</sub> equivalent emissions than chicken meat.

Veganism appeals because it prevents most harm, and because it is consistent where ethical vegetarianism is not.

Can you tell me about your transition to veganism? How long did this take? How did you carry it out?

It took me three years to accept the seriousness of this and switch. I didn't know any vegans (even the aforementioned lecturer ate cheese). Milk was the biggest hurdle; I am very into all-day cereal. So I tried every form of plant milk

I could until I got used to it. (Almond and horchata win.) I still miss pizza sometimes, but I actually no longer think about my diet from day-to-day; it is second nature.

Were there any challenges or personal concerns you had during your transition? How did you deal with them?

I had to cook properly for the first time in my life. And my decision was and is mocked in friendly terms by my high school friends (but never university ones). I researched the nutrition quite intensely and still keep up a solid regime. I take B12, vitamin D, creatine, and choline, to cover potential deficiencies. (These “subclinical” deficiencies aren’t so well studied, but the remedies are safe and cost less than £1 per day, so it’s a good deal.) The odd potential effects of soy isoflavones on hormonal balance was another one; I eat below 25g of raw soy per day.

Health, as a standalone motive for veganism, is not well-supported; the studies that show e.g. decreased cancer or cardiovascular disease are selecting from a population that is already unusually health conscious (and I don’t know of any proper controlled trials). It could be true, but at present it isn’t warranted. I want more people to go vegan, but decisions should be evidence-based or go home.

Are there any challenges or concerns that you have today with regards to your veganism? How does this make you feel?

None really. I live in Glasgow and London which are both amazing for it.

You stopped being vegan for a short period of time: what were your reasons for this? What made you return to veganism? How did stopping make you feel?

I lived in Tanzania in 2012; the available vegan food consisted of plain haricots, spinach, cassava and potato; not at all complete enough, in protein terms, for a long-term diet. The family had a well-treated cow, so I milked it and had boiled milk with breakfast. I was completely fine with this decision, until I learned that I had contracted giardiasis, probably from that milk!

Has your sex and/or gender ever been brought up as an issue/subject with regards to your veganism in any way? How did this make you feel?

It was an issue when I lived in China, where meat still has a status that it has largely lost here; many men eat as much meat as they can, for both class and gender signalling. I was teased for being squeamish or feminine in both the UK and China, but more in China. (In Tanzania it was sometimes respected as very sophisticated, but never emulated.)

I don’t mind; my gender is not really relevant to me (except insofar as being male and not having dysphoria has probably made my life easier). I don’t understand people who are stung by the above kind of mockery.

Do you promote veganism in any form? If so, in what ways? What has the reaction been to this?

I don't actively promote it; I view my role as normalising the practice by not being single-minded or stereotypical about it. I lie in wait; people usually bring it up themselves and the subsequent conversations are my contribution. I have turned perhaps a dozen friends onto the necessity of it. I don't have formal research to back this up, but I suspect that this method prevents fruitless interactions caused by vegan dogmatism and omnivores' "anticipated reproach".

I recognise the need for louder activists; I will donate to the Humane League, a transparent, evidence-based animal organisation (with an amazing name) who do this work well. I would support the criminalisation of factory farms, if that turned out to be a more effective way of solving the problem than e.g. alternatives like in vitro meat or (in the short-term) meat offsets.

You chose to be vegan for reasons that could be seen as trying to work towards a much bigger cause. Do you feel this way, like that you are part of a much larger movement? Or do you practice veganism and your choices are purely a personal thing?

Yes; I am an abolitionist about involuntary suffering; I have crazy sci-fi views about how we might achieve this.

The natural world seems an appalling place; billions upon billions of creatures starving or being eaten alive or raped every day. But we simply do not have the capacity to do much for them now; ecology is far too complex for us to know that intervention would not cause even more harm. There is a fledgling academic literature on the topic, but we are a long way from fixing this.

The least we could do is not make the problem worse, but people inexplicably support increasing the number of obligate murderers in the world (that is, "reintroducing wolves" and all that).

We are fortunate that science and industry are developing a way for us to end this quickly without unrealistic social change or a potentially counter-productive legal ban on factory farms: cultured meat (physically identical animal protein, produced without harm through cell biology) is here, and its price has dropped by a factor of ten thousand in a few years; some (biased) people forecast that mass produced cultured meat could undercut factory farms and drive them out of business within 30 years.

This may be already covered, but: how do you set boundaries for yourself on what is and is not acceptable to buy or consume? And are there any situations in which you would be more lenient?

The general principle is: no unnecessary harm to anything that can feel it. In modern urban life, this means: don't go hunting and don't buy anything any way that gives money to the harmful industries. In traditional societies where hunting is still a primary food source, it means: hunt with large guns, not bows.

In poor mountainous societies where sufficient crops cannot be grown it means keeping goats and sheep is permissible (until the world trade network links up with you and offers a fair price for soya).

My friend has chronic anaemia: her eating the occasional fish is arguably necessary harm.

You can consume animal products without moral problems if e.g. they're taken from out the bins of supermarkets (actual theft is an economic signal however so none of that); if they're heirlooms; or if it was already dead (e.g. roadkill); or if we get more info about the combination part of the binding problem which lets us classify borderline cases.

(There is a chance that eating clams is morally neutral; they have no central nervous system, i.e. nowhere that signals could be integrated into an experience.)

I don't do any of the above, for aesthetic reasons (as well as because my protease levels have changed so much in 10 years that it would probably be a very unpleasant experience!)

You're a member of an organisation, Giving What We Can, that helps charities proven to deliver much needed care at a low cost. With regards to your own ethical framework, is it the case, that this is more important than being a member of an organisation that is more focused on animals rights or promoting veganism? What are your reasons?

GWWC is actually closely associated with the work of Peter Singer, perhaps the most famous animal rights thinker. But it's true that they prioritise the suffering of humans - but this isn't necessarily an ideological decision, since we have a principled (and partially objective) way of ranking which organisations to support: the QALY per dollar. I donate to the Humane League (animals), the Against Malaria Foundation (humans), and GiveWell (incredibly deep research into charity effectiveness, including animal charities), in that order. If any animal organisation shows itself to be more effective (measured in QALYs per pound) than the first two then I will switch to them too.

Is being a member of an organisation that promotes veganism an important part of your ethics?

Not inherently; only insofar as the meat industry is the worst thing in the world and insofar as the ways that one person can tackle it are as powerful as the ways I can tackle e.g. malaria in humans. I am very pessimistic about collective action in this case; most people simply do not care about meat animals, and will not switch until we make cultured meat cheaper than factory meat.

I'm not a joiner really; I only joined GWWC in order to commit myself to a nonselfish life. No way out now, not without looking like a dick!

## **See also**

- Ethics > climate in a small sample of Scottish vegans

# Automatic for the people

Gavin

2016-10-20

{% include nonvicious/links.md %}

Factories that run ‘lights out’ are fully automated and require no human presence on-site... these factories can be run with the lights off.

- [Wiki](#)

Not only is it lights-out - we turn off the air conditioning and heat too.

- [an executive at Fuji Automatic Numerical Control](#)

Autonomous trucks are now in use and are already safer and more fuel-efficient than human driven ones. Truck drivers are 2% of the entire American workforce.

Crap journalism (that is, 80% of journalism) is now fully automatable. Automatic art is quite good and improving fast. Consider also the cocktail bartender. And so on: around half of all jobs are at risk of being automated, assuming the rate of AI progress just stays constant (“over an unspecified period, perhaps a decade or two”).

```
<h3>Types of automation</h3>
<div>
    {% include nonvicious/types.html %}
</div>
```

Automation has been happening for hundreds of years, but in the past it probably didn’t produce long-term or “technological unemployment”. This is probably because people were easily able to think up new professions given new tech and culture, and since the increased productivity translated into lower costs for the automated good, which stimulated other parts of the economy, and people could retrain for the subsequent new jobs. The present wave might be different: it might actually reduce the number of available jobs permanently, because the machines now entering the workforce can be applied to many of the jobs that people could retrain to do.

If so, our economy - resource allocation based on employment (which we use as a poor proxy variable for productivity) - is a local maximum and we cannot expect to arrive at a good outcome without activism, since:

The new machine-learning automation could fully replace around half of jobs.

Since these jobs involve even our highest cognitive faculties, it is possible that we won't think up new productive jobs for the replaced workers, like we did in the past.

So automation could produce an unprecedentedly high unemployment rate, ~60%.

Most existing unemployment welfare systems are inadequate and degrading.

So without intervention, a crash in human welfare is easily possible.

But, unless we automate a lot more, we the species will never have enough wealth to offer a decent basic income, and everyone will continue to waste half their lives at work. Like C20th peasants.

```
<h3>How much is there for everyone?</h3>
<div>
{%
  include nonvicious/each.html
}
</div>
```

Automation is maybe the main way that technology improves most people's lives: aside from foolish status exceptions like Apple products, big reductions in manufacturing cost usually mean big reduction in the end cost of goods. Obviously, replacing labour costs with lower-marginal-cost machines benefits rich machine-owners most, but automation also allows giant price cuts in all kinds of things; over the last two centuries these cuts have transformed society, increasing equality enormously by making things affordable for the first time.

Besides the obvious example - that we now produce a volume of food far beyond the needs of the entire world population (2940kcal per person per day, though with terrible distributive failures) - consider that a single ordinary shirt takes 508 hours of labour to produce on a spinning wheel and hand-loom - so you would expect to pay something above \$3600 at current minimum wage (still \$900 at 1400CE wage levels). 4 Getting costs as near to zero as possible is the way we will solve the easy problem of human existence, scarcity of basic goods and leisure. 5

```
{%
  include nonvicious/solutions.html
}
```

I am not very sure of any of the above; the actual stats on productivity growth are worrying for the opposite reason: it has been too slow to support wages for a long time. Anyway other powerful forces (e.g. global outsourcing, the decay of unions) besides robots have led to the 40-year decline in labour's share of global income. But those will produce similar dystopian problems if the trend continues, and there's enough of a risk of the above scenario for us to put a lot of thought and effort into protecting people, either way.

(c) GE Thyer (1988)

```
{%
  include nonvicious/foots.html
}
```

# The Worst Game Ever

Gavin

2016-10-21

```
{% assign supe = "https://en.wikipedia.org/wiki/Superrationality" %} {%  
assign gr = "https://en.wikipedia.org/wiki/Grim_trigger" %} {% assign  
noi = "http://www.cs.utexas.edu/~chiu/papers/Au06NoisyIPD.pdf" %} {%  
assign pd = "https://en.wikipedia.org/wiki/Prisoner's_dilemma" %} {%  
assign tit = "https://en.wikipedia.org/wiki/Tit_for_tat#Game_theory" %}  
{% assign mic = "https://learn.saylor.org/course/ECON101" %} {% assign  
kin = "https://en.wikipedia.org/wiki/Kin_selection" %} {% assign model  
= "http://web.archive.org/web/20200707011614/http://www.bostonfed.org/-/  
/media/Documents/neer/neer697b.pdf" %} {% assign nash = "http://web.cse.ohio-  
state.edu/~stiff.4/cse3521/prisoners-dilemma.html" %} {% assign king =  
"https://unherd.com/2020/02/the-madness-of-mervyn-kings-uncertainty" %}  
{% assign berg = "https://link.springer.com/chapter/10.1007/978-1-349-20181-  
5_26" %} {% assign rpw = "https://umass.app.box.com/s/n72u3p7pyj/folder/80549398"  
%}
```

I was at a corporate team-building event, because I wasn't persuasive enough to not be. I was a prisoner.

The organisers set up a game: a three-player, unknown-length iterated Prisoner's dilemma. There was no initial discussion, but free discussion every two rounds. Payoffs were the standard unitless numbers, shifted so that some outcomes were negative. Scores began at zero. No objectives were given.

All co-op

Some defect

All defect

All +1

Defectors +2Co-op: -1

All -2

I *swear* I am not making up the roles the players tacitly settled into: perfect archetypes of game theory.

a Homo economicus (who clearly took Micro 101 and nothing further).1

an ineffective altruist, trying to get everyone to the global maximum, but without any leverage or provocability. Opened with co-operate, tried again after every negotiation round until the second half when he got in a huff.

a noise generator. Random action, or, action based on his reading of opponents' body language.

A:

It's totally straightforward, there's only one right answer: for Prisoner's dilemmas, the only Nash equilibrium is 'always defect', because it only makes sense to defect in the final round, and the inference to prior rounds is timeless.

Me:

No, that's for the known-length case with two players! And for when you can assume perfectly rational opponents! And for when their actions are independent of yours, with no communication! And you're aiming for personal loss minimisation, but you weren't given a loss function: you don't know that negative scores are non-fatal (or that they allow for "losing the least").

He didn't listen. Headless in the grip of theory, the ecstasy of proof, (A) defected every time. During the communication, he was sometimes honest about his strategy and sometimes pretended to accept a truce.

The game ran about 15 rounds, I think ending early in despair. We ended with all scores negative, between 1 and 10 points under. It was announced that everyone had lost, since -1 represented death. Everyone sulked, especially the organisers.

This was in fact an *incredible* lesson, but not the one the organisers wanted me to learn.

- Decisions make no sense without a loss function! Probably the organisers had no idea about loss functions or equilibria or mathematical induction backwards in time, but I could have drawn it out of them. (Scary thought: society depends on statistical inference, and yet some massive majority of those inferences (the null-hypothesis significance tests) are made in total ignorance of their implicit decision theory.)
- What's wrong with co-operating every time? Well, setting aside the poor bot's own (generally terrible) outcome: it invites exploitation, and so can actually be destabilising in a sense, compared to precommitted tit-for-tat.
- It is really amazing how stupidly a clever person can act if they are relying on a clever false theory. This can result from any method, any species of reasoning, but using maths badly is the most complete way of disabling such a person. This is why, despite appearances, we have to listen to mouthy gits like Taleb: there really are model error monsters out there.

- Even in the absence of A's model error, the presence of noise would have completely destabilised the equilibrium anyway. We were doubly doomed.
- The game wasn't long enough for the other key ingredient of super-rationality to arise: forgiveness, necessary in all closed-source stochastic domains, like life.
- Body language reading *could* be a real skill, but either way I think most people don't have enough skill to substitute for outside-view reasoning - even in toy examples like the above.

### **Anthropics and game theory only work on themselves**

When reasoning about what I should do, if I reflect on what my predecessors must have done in order for me to exist, and then generalise this to my descendants (since they are a sort of partial copy of me), I could convince myself that I should do what would maximise my chance of coming into being. (This is all under the assumption that I am the sort of being which should exist, at least equally compared to the counterfactual people in other chains of descent.) But this doesn't work unless the other people in the chain are also doing anthropics.

Similarly, you don't benefit from doing game theory on an unpredictable opponent (for instance one who doesn't know or rejects game theory).

### **See also**

- *The Use and Abuse of Formal Models in Political Philosophy*, Robert Paul Wolff.

*Thanks to Misha Yagudin for the anthropics point.*

Not a real Homo economicus, obviously: instead a Homo sapiens running a bad simulation of one, at the same time shutting down the common sense that might have saved him.

# Effective Altruism Global: x: Oxford

Gavin

2016-11-21

```

{% assign mood = "https://lukemuehlhauser.com/musks-non-missing-mood/" %}
{% assign nate = "https://forum.effectivealtruism.org/posts/hkimyETEo76hJ6NpW/on-
caring" %}
{% assign ocb = "https://www.youtube.com/watch?v=67oL0ANDh5Y#action=share" %}
{% assign parfit = "https://www.youtube.com/watch?v=KtU0pah4R8Q" %}
{% assign anders = "https://www.youtube.com/watch?v=kt_8nLrUkI8" %}
{}
```

I'm not a joiner. 1 But I have a lot of strange ideas, and a lot of odd energy, and a lot of unusual feelings, and these usually mislead people who go off on their own. So it's a stroke of incredible fortune that a movement of people with these things happens to arise - just as I graduate and try to become technical enough to understand what the best thing to do is.

I'm not sure I've ever experienced this level of background understanding, these tiny inferential distances, in a large group. Deep context - years of realisations - mutually taken for granted; and so shortcuts and quicksteps to the frontier of common knowledge. In none of these rooms was I remotely the smartest person. An incredible feeling: you want to start lifting much heavier things as soon as possible.

One liners:

Effective altruism is to the pursuit of the good as science is to the pursuit of the truth.

(Toby Ord)

If the richest gave just the interest on their wealth for a year they could double the income of the poorest billion.

(Will MacAskill)

If you use a computer the size of the sun to beat a human at chess, either you are confused about programming or chess.

(Nate Soares)

Evolution optimised very, very hard for one goal - genetic fitness - and produced an AGI with a very different goal: roughly, fun.

(Nate Soares)

The goodness of outcomes cannot depend on other possible outcomes. You're thinking of optimality.

(Derek Parfit)

---

  
Soares, Ord, Krakovna, Shanahan, Hassabis, MacAulay.

---

## Prospecting for Gold



Owen Cotton-Barratt formally restated the key EA idea: that importance has a highly heavy-tailed distribution. This is a generalisation from the GiveWell/OpenPhil research programme, which dismisses (ahem, “fails to recommend”) almost everyone because a handful of organisations are thousands of times more efficient at harvesting importance (in the form of unmalarial children or untortured pigs or an unended world).

Then, Sandberg’s big talk on power laws generalised on Cotton-Barratt’s, by claiming to find the mechanism which generates that importance distribution (roughly: “many morally important things in the world, from disease to natural disasters to info breaches to democides all fall under a single power-law-outputting distribution”).

Cotton-Barratt then formalised the Impact-Tractability-Neglectedness model, as a piece of a full quantitative model of cause prioritisation.



Then, Stefan Schubert’s talk on the younger-sibling fallacy attempted to extend said ITN model with a fourth key factor: awareness of likely herding behaviour and market distortions (or “diachronic reflexivity”).

There will come a time - probably now - when the ITN model will have to split in two: into one rigorous model with nonlinearities and market dynamism, and a heuristic version. (The latter won’t need to foreground dynamical concerns unless you are 1) incredibly influential or 2) incredibly influenceable in the same direction as everyone else. Contrarianism ftw.)

---

What is the comparative advantage of us 2016 people, relative to future do-gooders?

Anything happening soon. (AI risk)

Anything with a positive multiplier. (schistosomiasis, malaria, cause-building)

Anything that is hurting now. (meat industry)

---

### Sandberg: one-man conference



Anders Sandberg contributed to *six* events, sprinkling the whole thing with his hyper-literate, uncliched themes. People persisted in asking him things on the order of “whether GTA characters are morally relevant yet”. But even these he handled with rigorous levity.

My favourite was his take on the possible value space of later humans: “chimps like bananas and sex. Humans like bananas, and sex, and philosophy and competitive sport. There is a part of value space completely invisible to the chimp. So it is likely that there is this other thing, which is like whooooaa to the posthuman, but which we do not see the value in.”

---

- Books usually say that “modern aid” started in ‘49, when Truman announced a secular international development programme. Really liked Alena Stern’s rebuke to this, pointing out that the field didn’t even try to be scientific until the mid-90s, and did a correspondingly low amount of good, health aside. It didn’t deserve the word, and mostly still doesn’t.
  - Nate Soares is an excellent public communicator: he broadcasts seriousness without pretension, strong weird claims without arrogance. What a catch.
  - Dinner with Wiblin. My partner noted that I looked flushed. I mean, I was eating jalfrezi.
  - Catherine Rhodes’ biorisk talk made me update in the worst direction: I came away convinced that biorisk is both extremely neglected and extremely intractable to anyone outside the international bureaucracy / national security / life sciences clique. Also that “we have no surge capacity in healthcare. The NHS runs at 98% of max on an ordinary day.” This harsh blow was mollified a bit by news of Microsoft’s mosquito-hunting drones (for cheap and large-sample disease monitoring, not revenge).
- 

### Inequality vs impact

Most sessions I attended had someone asking the same desultory question: “how might this affect inequality?” (AI, human augmentation, cause prioritisation

as a priority.) The answer's always the same: if it can be automated and mass-produced with the usual industrial speed, it won't. If it can't, it will.

Actually it was good to ask (and ask, and ask) this for an ulterior reason:

Molly Crockett's research - how a majority of people might relatively dislike utilitarians - was great and sad. Concrete proposals though: people distrust people who don't appear morally conflicted, who use physical harm for greater good, or more generally who use people as a means. So express confusion and regret, support autonomy whenever the harms aren't too massive to ignore, and put extra effort into maintaining relationships.

These are pretty superficial. Which is good news: we can still do the right thing (and profess the right thing), we just have to present it better.

(That said, the observed effects on trust weren't that large: about 20%, stable across various measures of trust.)

```

```

---

## The Last Dance of Derek Parfit

```

```

Very big difference in style and method between Parfit's talk and basically all the others. This led to a sadly fruitless Q&A, people talking past each other by bad choice of examples. Still riveting: emphatic and authoritative though hunched over with age. Big gash on his face from a fall. A wonderful performance. Last of His Kind.

Parfit handled 'the nonidentity problem' (how can we explain the wrongness of situations involving merely potential people? Why is it bad for a species to cease procreating?) and 'the triviality problem' (how exactly do tiny harms committed by a huge aggregate of people combine to form wrongness? Why is it wrong to discount one's own carbon emissions when considering the misery of future lives?).

```

```

He proceeded in the (IC20th) classic mode: state clean principles that summarise an opposing view, and then find devastating counterexamples to them. All well and good as far as it goes. But the new principles he sets upon the rubble - unpublished so far - are sure to have their own counterexamples in production by the grad mill.

The audience struggled through the fairly short deductive chains, possibly just out of unfamiliarity with philosophy's unlikely apodicticity. They couldn't parse it fast enough to answer a yes/no poll at the end. ("Are you convinced of the *non-difference view*?"

The Q&A questions all had a good core, but none hit home for various reasons:

Does your theory imply that it is acceptable to torture one person to prevent a billion people getting a speck in their eye?

Parfit didn't bite, simply noting, correctly, that 1) Dostoevsky said this in a more manipulative way, and 2) it is irrelevant to the Triviality Problem as he stated it. (This rebuffing did not appear to be a clever PR decision - though it was, since he is indeed a totalarian.)

Sandberg: What implications does this have for software design?

Initial response was just a frowning stare. (Sandberg meant: lost time is clearly a harm; thus the designers of mass-market products are responsible for thousands of years of life when they fail to optimise away even 1 second delays.)

I'd rather give one person a year of life than a million people one second. Isn't continuity important in experiencing value?

This person's point was that Parfit was assuming the linearity of marginal life without justification, but this good point got lost in the forum. Parfit replied simply - as if the questioner was making a simple mistake: "These things add up". I disagree with the questioner about any such extreme nonlinearity - they may be allowing the narrative salience of a single life to distract them from the sheer scale of the number of recipients in the other case - but it's certainly worth asking.

We owe Parfit a lot. His emphasis on total impartiality, the counterintuitive additivity of the good, and most of all his attempted cleaving of old, fossilised disagreements to get to the co-operative core of diverse viewpoints: all of these shine throughout EA. I don't know if that's coincidental rather than formative debt.

(Other bits are not core to EA but are still indispensable for anyone trying to be a consistent, non-repugnant consequentialist: e.g. thinking in terms of degrees of personhood, and what he calls "lexical superiority" for some reason (it is two-level consequentialism).)

The discourse has diverged from non-probabilistic apriorism, also known as philosophy, the Great Conversation. Sandberg is the new kind of philosopher: a scientific mind, procuring probabilities, but also unable to restrain creativity/speculation because of the heavy, heavy tails here and just around the corner.



Incredibly beautiful setting (Exam School). Incredibly professionally organised by undergraduates, chiefly Oliver Habryka and Ben Pace.

{% include eagox/foots.md %}

# Estimating political controversy

Gavin

2017-01-07

```
{% include controversy/links.md %}  
{% include controversy/controversy-simple.html %}  
{% include controversy/controversy-technical.html %}
```

# The presumed worth of anthropology

Gavin

2017-01-11

{% include anthro/links.md %}

Olduvai Gorge (2009), by Noel Feans

Social science is *hard*. Most social phenomena involve [thousands][Causa] [of subtle effects][Manzi], of several [different][Proph] [ontological sorts][Reason], all acting on relevant populations up to about [a billion units][Nation] in size - and even if you were ever able to cement any results about it, and publicise them, a portion of your subjects would immediately [change in response][Reflex] - to spite you.

But it's important and noble to try, anyway, to understand human things using our best toolkit.

However, some anthropologists I've talked to reclaim the title of 'science' entirely - as, indeed, did the [American Anthropological Association in 2010][Wade]:

Until now, the association's long-range plan was "to advance anthropology as the science that studies humankind in all its aspects." The executive board revised this last month to say, "The purposes of the association shall be to advance public understanding of humankind in all its aspects." This is followed by a list of anthropological sub-disciplines that includes political research. The word "science" has been excised from two other places in the revised statement...

1

To put it lightly, this is counterproductive.

1. First of all, it is a straightforwardly false claim about their peers - physical anthropologists and linguistic anthropologists have made [great][Kurgan] [and][Bones] [enduring][Homin] scientific discoveries. So this hegemonic statement is really a claim about *cultural* anthropology, a subfield we can usually paraphrase as literary anthropology. 2
2. It treats science as something alien to ordinary life, something distinctively Western. However, the only thing uniting the sciences is *taking evidence as seriously as possible*: pragmatically, 'science' is the practice of making systematic observations, relating these to possible explanations, and offering

data and argument to peers for criticism and replication or disconfirmation. Ethnographers are absolutely scientists in this broad sense, even though some of their judgments (for instance about the semantic and symbolic aspects of a culture) have not been amenable to computational or statistical analysis. 3 And it is absurd to suggest that non-Westerners have not played large parts in the endeavour, fast increasing.

3. Eliminative reductionism, excessive formalism and political domination are nowhere implied by the scientific method. As a result, taking pride in not ‘doing science’ is chilling to me. 9 Yes: they’re trying to be humane. Yes, they’re trying to resist [eliminative reductionism][ElMat], [mathiness][Mathy], and racism (the “arrogant perceptions of the weaker peoples as instrumental means of the global projects of the stronger”). 4 But in the process they attack our only hope of ever getting above divisive tribal instincts, or controlling for self-interest, chance, delusion and whim. Any political movement which does not take epistemic care will sooner or later do harm, and possibly more harm than good.

```
<h3>Cognitive imperialism</h3>
<div>Some people, like <a href="{{Batshitste}}>Marie Battiste</a>, call the discovery and
<blockquote>
  <i>Cognitive imperialism is a form of cognitive manipulation used to disclaim other know
</blockquote>
Objective truth <i>is</i> imperialist in some sense – but it wasn't scientists that made the
<br><br>
The idea that politicised and relativised epistemology has <i>helped</i> (or would help) the
<blockquote>
  <i>The [Bush] aide said that guys like me were "in what we call the reality-based commun
</blockquote>
What's worse than being abducted, tortured and your family never hearing of you again? Surel
</div>
```

---

Unscientific research (where scientific means could be fruitful) is just discreditable: it is easily and often the exquisite interpretation of statistical noise. Unfortunately, the unscientific manner is here paired with active disdain of those cultural anthropologists who do attempt science: e.g. [Jared Diamond][Diamond], [Napoleon Chagnon][Chagnon], [Steven Pinker][Survive], Pascal Boyer. 5 Scientific cultural anthropology is a serious, interdisciplinary research programme shared with genetics and psychology. (It can be googled via “[cultural evolution][CultEv] [theory][DualInt]” or “coevolution”.). Given that I haven’t read much of it, given that I value general truth more than particular conjecture, why would I spend time on the non-scientific kind?

```
<h3>Mob science</h3>
<div>
  I continue to struggle to distinguish anthropology and <a href="{{Phenom}}>soft</a> soc
```

</div>

The first problem, beyond any of the totally speculative ideological effects, is its opportunity cost: people tend to wear only one hat, and wearing the cultural anthropology hat means they probably won't wear any of the fantastic, proven, world-changing hats.

---

Just because c-anthropology is very confused on *average* doesn't allow us to write off the whole field. 90% of everything is crap after all. 11 For instance: I only recently realised the distinctive value of the philosopher Rousseau, an anthropologist in spirit: few Europeans really considered that there could be value in foreign cultures in his time. 6

This sounds incredible in our hybridised culture, for we are obsessed with [fleeing ourselves][Travel], and have a ridiculously rose-tinted idea of what pastoral (read: foreign) life was and is actually like. But you can see disdainful total superiority in even the most intelligent commentators from the C17th, for instance Samuel Johnson. 7

If anthropology is *a* cause of the huge shift in attitudes from then until now, then they will have done us a great service. (For all that it is anthropologists who now lead the quest for segregation and cultural purity, in the form of attacking often-harmless cultural exchange as cultural appropriation.)

So, in spite of them, I'm going to find the best of the field, to “[steel-man][Steel]” them. Out of charity, I won't cover the kind of writer who believes in the [equipollence][Equi] of ways of knowing.

```
<h3>Research programme</h3>
<div>
  I'm being combative, but I am actually interested. I'm looking for answers to: <br><br>
  <ul>
    <li>What has cultural anthropology achieved?</li>
    <li>How much of the modern love of foreign experience and culture can be attributed to t
    <li>Is cultural anthropology unusually ideologically polluted?</li>
    <li>How does cultural or social anthropology differ from sociology?</li>
    <li>What is cultural anthropology's method?</li>
    <li>Does cultural evolution theory subsume or supercede cultural anthropology?</li>
    <li>How unscientific is cultural anthropology? How can we tell?</li>
  </ul>
</div>

<h3>e.g.: <i>What is the anthropological method?</i></h3>
<div>I don't really know and neither of the textbooks I tried were very illuminating. Here's
<ul>
  <li><i>Methodological relativism</i>. The ethnographer must put aside the literal truth
  <li><i>Lionisation of culture</i> as the indispensable framework for human life (they ar
  <li><i>Universalisation of culture</i>. Everything (every datum, every theory, every exp
```

```

<li><i>Advocacy</i>. That you can't just _describe_ a dying way of life, you have to make it happen</li>
</div>

<h3>e.g.: <i>How unscientific is cultural anthropology?</i></h3>
<div>Some ways we could investigate this:<br><br>
<ul>
    <li>Email everyone listed on every department page and ask. I can imagine impressive responses</li>
    <li>Gather up all public responses to <a href="{{Fail}}">#AAAFail</a> and see what % were positive</li>
    <li>Crawl JSTOR and look for signs of quantitative reasoning in papers (numerals, math notation)</li>
</ul>

My survey would probably just ask a Likert question for each variable in <a href="{{earlyModern}}>Early Modern</a>
<ul>
    <li>Are their explanations couched only in terms of natural phenomena? </li>
    <li>Is their research based on going and looking at things?</li>
    <li>Is their method quantitative where appropriate?</li>
    <li>Is their method reductionist?</li>
    <li>Are they fallibilist about their conclusions?</li>
    <li>Do they share their data and invite criticism and replication?</li>
</ul><br>
    (Obviously many self-described scientists fall short of these, for good and bad reasons)
</div>

<h3>Reading List</h3>
<div id="listFrame">I couldn't find a reading list for "the cultural anthropology canon". The best I found was this</div>
<% include js/lazyFrame.html %>
<script>
    var src = "https://docs.google.com/spreadsheets/d/1EuFXFPpzRCG9Vjsb8zYPMmAPAnqM2bd_UUg/edit#gid=1";
    definiteEvent( createIframe, [src, "listFrame"] );
</script>
</div>

<% include anthro/foots.html %>

```

# Ludwig and the Machine

Gavin

2019-01-07

Howson on Bayes

Cox on induction

Solomonoff's razor

Guaranteed induction

Parsimony refuted and rehabilitated

The logical analysis of concepts without use of necessities and sufficiencies.

Ensemble success vs Strong Occam's razor

Ryle's knowing-that (GOFAT) and knowing-how (CLT, SLT)

Here is a toy model of aesthetics with just two binary variables, 'classiness' and 'busyness' 1:

Minimalism: Simple Classy

Baroque: Busy Classy

Brutalism: Simple Vulgar

Rococo: Busy Vulgar

Are these descriptions true? Well, they are incomplete, and are not definitions (i.e. one-to-one mappings), but yes. Are they helpful? As a start, absolutely.

Now, the labels on the left are vague and intuitive family resemblances; it is a fool's game to imagine they could ever be nailed down as monothetic definitions (the philosopher's ideal of neat, necessary and sufficient sets of attributes). We can still model usefully and harmlessly, even if the models can never be complete.

2

But the critics and art academics I know spend far more time muddying the water: deconstructing our use of the problematic term "classy"; and who gets to say what 'simplicity' is anyway? They don't seem to want to explain things, even fuzzily.<sup>3</sup> Or, maybe they do, but refuse to accept anything but a perfect final omniperspectival explanation (the like that can never be supplied), maybe to keep themselves in work.

Imagine if critics were conscientious enough to build a consistent hundred-variable, real-valued theory of art. Would it “solve” criticism? Never ever. Would it make the points of disagreement between interpretations more vivid? Would it force clarity in this, the most pompous and vacuous discourse? Yes.

But we will probably have to wait for AI art critics for that, to go with the excellent AI artists we have already.

There's no fixed criteria for these terms, you say? There's too much political context and social problematics involved for art to be tackled by statistical inference, you reckon? Well, machine learning is the automatic empirical discovery of non-necessary, non-sufficient attributes; it can and will cover the full range of the term's application and will do so by frequency, not political agenda.

The polythetic wall held up against philosophers and computers for a long time, sixty years at least. But it's time. Wittgenstein:

someone might object: “You... have nowhere said what the essence of... language is: what is common to all these activities, and what makes them into language or parts of language. So you let yourself off the very part of the investigation that once gave you yourself most headache, the part about the general form of propositions and of language.”

And this is true. — Instead of producing something common to all that we call language, I am saying that these phenomena have no one thing in common which makes us use the same word for all, — but that they are related to one another in many different ways. And it is because of this relationship, or these relationships, that we call them all “language”. I will try to explain this.

The result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. // I can think of no better expression to characterize these similarities than ‘family resemblances’...

Basically, Wittgenstein is pointing out that the classic philosophical approach to conceptual analysis (the “monothetic” approach) fails, because real concepts are messy and defined in a partial-membership way, (“polythetic”). He implies that this can't be , which was a fair enough point in 1950.

But then Nils Nilsson:

Some tasks cannot be defined well except by example; that is, we are able to specify input/output pairs but not a concise relationship between inputs and desired outputs... machines [are now] able to adjust their internal structure to produce correct outputs for a large number of sample inputs and thus suitably approximate the relationship implicit in the examples.

4

We can do this to anything we have at least proxy data for, which is, arguably, every thing that could matter. (If you count e.g. self-report of experience as

reliable proxy data for consciousness.)

This is far from the most important way ML affects old thought.

The formal sciences, math/stats/CS/decision theory.

- inductive bias; the set of assumptions a learner uses to predict outputs given inputs that it has not encountered *Absolute bias: constraint on hypothesis space. e.g. search only linearly separable functions* Preference bias: select the optimal hypothesis according to some ordering scheme. e.g. least Kolmogorov complexity
- statistical bias: directional error in an estimator. error you cannot correct by repeating the experiment many times and averaging together the results.
- cognitive bias:

Guarantees rarely have practical relevance: you're likely to have benchmarked and amortized a hundred thousand runs by the time the theoretician has thought up a proof for what you've already seen. And even if you have a guarantee before starting, your benchmarks will tell you far more about the system's actual usefulness - the guarantee tends to be a ridiculous underestimate. But proof is a fine thing even so, and it is on this level, the absolute apriori, that most philosophy thinks to live.

You don't know how right or wrong it will end up being - but you do know that it won't be worse than [Bound]. You can't guarantee that it'll settle down in your lifetime, but you can guarantee that the probability of it never settling down is low.

## Platonic forms

Olah and Plato

<https://distill.pub/2018/building-blocks/>

This is a compressed essence of dog

## Problem of induction

Major development of recent decades has been guarantees on the results of induction, particularly if we're willing to settle for probabilistic guarantees.

"this was perhaps Hume's first great discovery... What he finds is that the confidence we have in natural law — in the regularities and uniformity of nature, in the future being about to resemble the past — has a source in our animal nature. Animals too expect things to go on much as they have gone on — but it has no justification in reason. There is no a priori way of showing that it's even probable that the future will resemble the past... There's nothing available to our understanding to show us why things must keep on as they apparently always have." <https://fivebooks.com/best-books/david-hume-simon-blackburn/> Computational Learning Theory and radical scepticism What justifies belief in

an external world? != When can we expect a hypothesis to predict future data?  
CLT gives

Tradeoff between hypotheses considered and prediction confidence \* Occam's razor theorem Upper bound for data required for confidence Relation of training and test performance PAC: Even in the worst-case, we can learn. We don't need to know the evaluation distribution (the truth) to approximate it. Given m examples ( $X$ ) from D and labels ( $f(x)$ ), find a hypothesis  $h$  such that  $P(h(x) == f(x)) > 1 - \epsilon$  (Alternative to Bayesian inference? Don't need a distribution over hypotheses: some distribution over sample data is enough)

## Causality.

```
{% assign gabgoh = "https://gabgoh.github.io/ThoughtVectors/" %}
```

A solid addition for a three-variable version would be "flat or shadowy" i.e. using clean planes or chiaroscuro. This would let us introduce Classical (simple classy flat) and Gothic (busy classy shadowy) and three others I can't be bothered looking up. Though this is a distinctively visual epithet, where the above should apply to all arts.

What are the risks of building a model? Does a model obscure reality behind its necessarily limited representation? No; all the authors and users of models need, to avoid delusion and harm, is a little imagination and humility.

There are of course honourable exceptions.

See also the impressive but inexplicable-in-a-given-instance behaviour of feedforward neural nets:

Neural networks have the rather uncanny knack for turning meaning into numbers. Data flows from the input to the output, getting pushed through a series of transformations which process the data into increasingly abstruse vectors of representations... But the vectors themselves have thus far defied interpretation.

- here

# Machine Learning Simply

Gavin

2017-01-18

# Analysis of all Kaggle-winning algorithms

Gavin

2017-02-10

Kaggle-winning Algorithms, 2010-2016

```
CompetitionId,  
Competition segment,  
End date (Deadline),  
# Rows,  
# Cols,  
EvaluationAlgorithm  
    Structured,  
    Winning team size,  
    Winning team # submissions,  
    Models used,  
    Ensemble?
```

These fields are taken directly from Meta Kaggle fields; bold are new, fields that I've collected myself.

how structured is this data?

If you had time, you could expand this analysis to the top 3 teams (who are required publish their scripts). If you had a lot of time, you could crawl the Kaggle forums and the wider web for entrants who post their scripts elsewhere (dozens per competition).

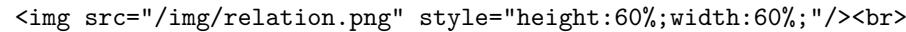
CTF.

# Data Science FAQ

Gavin

2017-01-28

```
{% include js/lazyFrame.html %} {% include dsfaq/links.md %}


<small>(c) Larry Leemis (2008) <a href="#fn:7" id="fnref:7">7</a></small>
```

Data Scientist: Person who is worse at statistics than any statistician  
& worse at software engineering than any software engineer.

~ Will Cukierski

What is data science?

```
{% include dsfaq/prozess.html %}
```

Modelling

```
<div class="accordion">
    {% include dsfaq/modelling.html %}
</div>
```

Machine learning

```
<div class="accordion">
    {% include dsfaq/ml.html %}
</div>
```

Programming

```
<div class="accordion">
    {% include dsfaq/code.html %}
</div>
```

Data janitoring

```
{% include dsfaq/janitor.html %}
```

Recommended reading

```
<div class="accordion">
    {% include dsfaq/books.html %}
</div>
```

<br><a href="{{spoilers}}>Full list here</a>. All books: I don't care for MOOCs or Youtube

```
<h3>Glossary</h3>
<div>
{%
  include dsfaq/gloss.html
%}
</div>
```

Inspired by this long list of things that separate a fresh programmer from an actual engineer. 2.

```
{%
  include dsfaq/foots.html
%}
```

# What's the highest moral wage?

Gavin

2014-01-11

```
{% include maxwage/links.md %}  
{%   include maxwage/max-simple.html %}  
{%   include maxwage/max-stylised.html    %}
```

# Wage theft and time theft

Gavin

2017-02-01

**clipping:** indirect reduction of employee wages **docking:** punitive reduction of pay. Bonded labour.

- More than contracted hours. (Salaried people: all overtime.)
- no travel expenses
- off the clock waiting or working (arrive early, skip a meal break, stick around after punching out)
- travel during plan hour doctoring a pay sheet. listed as independent contractors

Salaried (middle class) people are less

# Does the gut cross the epistemic barrier?

Gavin

2017-03-09

```
{% assign int = "https://plato.stanford.edu/entries/justep-intext/" %} {%  
assign dp = "https://www.hedweb.com/diarydav/2008.html" %} {% assign  
rs = "https://en.wikipedia.org/wiki/Radical_skepticism" %} {% assign prag  
= "https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1933-1592.2005.tb00507.x"  
%}
```

If there is a logical or epistemic barrier between the mind and nature, it not only prevents us from seeing out, it also blocks a view from the outside in.

– Donald Davidson

I worry that the closest I come to staying in touch with the real world is eating bits of it: the epistemology of food, so to speak.

– David Pearce

The ‘epistemic barrier’ is a thing between the mind and the external world: the thing that makes it possible to say that we do not have any knowledge.<sup>2</sup> It’s not very popular: there are dozens of arguments for why it isn’t there. (You can tell they didn’t work, because there are dozens of them and not one.)

Still, I (and the whole field) learned a lot about epistemology arguing about this stuff, and the thought still tickles me.<sup>3</sup> So here’s what I think Pearce was getting at:

The brain is made of food, ingested matter.

Knowledge inheres in the brain.<sup>1</sup>

So knowledge inheres in (metabolised) food.

Food, like all matter, is of the external world.

So the mind inheres in the external world.

So there is no metaphysical barrier between mind and world.

So there is no puzzle about the possibility of knowledge.

Clearly this does not defeat the radical sceptic in her original, Cartesian internalist problematic (“it’s an epistemic barrier, not a metaphysical one - I don’t grant

(1) or (2) or (4)"). But one good candidate for a philosophical fact is: nothing can. The only way to win is not to play.

Maybe not just in the brain, but that doesn't hurt the argument.

In the sophisticated Pyrrhonian form, "...not even of this sceptical proposition".

I don't think I'm a brain in a vat. But I'm vaguely annoyed by knowing that an actual brain in a vat would think exactly the same thing for the same reason.

— Scott Alexander

# History of applied machine learning

Gavin

2017-03-01

- Chess Economic impact:

Post office. Handwriting recognition (OCR) 1987: Algorithmic trading, finance, predicting stock ups and downs, starting in the late 1980s. 1995: Mining corporate databases Direct marketing, Customer relationship management, Credit scoring, Fraud detection. 2000: E-commerce: automated personalization quickly became de rigueur. Web search Ad placement. 2001: war on terror. 2005: social network analysis 2006: NLP for product buzz

molecular biologists astronomers

2012: Cars

# Paradigms are abstractions of abstractions

Gavin

2017-03-01

Peter Norvig lists five programming paradigms to learn, divided by their most distinctive forms of abstraction. This is good advice and well-conceptualised.

But these are hard to understand unless you already know them.

Two rephrasings to make this all fit under the abstraction banner:

- Declarative programming is implementation abstraction: you are instructing an engine to build a program for you that meets this spec.
- Parallelism is processor abstraction: split this task.

no abstraction. (binary machine code) statement abstraction (Assembly) procedural abstraction (Assembly) data abstraction (FORTRAN)

class abstraction (C#) functional abstraction (Haskell)

For type abstraction (Python)\*

syntactic abstraction (Scheme) implementation abstraction (Prolog) processor abstraction (Go)

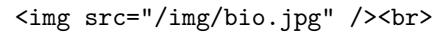
\* An extremely leaky abstraction, yes, but it's a leaky example of how it could work.

# Why care about genetic privacy?

Gavin

2017-03-28

```
{% include genes/links.md %}


Gene linkage in a small part of a genome
```

I'm keen to get my genome sequenced. This is so I can dig around in my data, snooping for clues about genealogy, psychology, and health. It's a highly novel kind of narcissism.

I'm mostly just waiting for the cost to plummet some more. One reason it isn't cheap yet is regulation. Most of this concerns data privacy, including my right to be protected from misinterpreting information I ask for and pay for.

I don't necessarily understand my society's extreme caution in this manner: I harbour an ambition to put my genome up online as a joke: so I can say "I am open source". But some people have already done this for serious reasons, to find unknown relatives.

Here are some things you might worry about if you did such a thing.

```
<h3>ACTUAL THREATS</h3>
<div>
  <ul>
    <li>
      <span class="b">Health insurance</span> <br>
      As a European, I found it hard to believe how callous American healthcare is. That's
      <i>Risk applies to</i>: People with known pathological alleles living in non-a
      <i>Risk to me</i>: Low.</li>
      <br>
    <li>
      <span class="b">Adversarial dirt.</span>
      <br>
      If you are ever in a position where someone needs to make you look bad - e.g. a
      <i>Risk applies to</i>: Extraordinarily ambitious people with known pathological
      <i>Risk to me</i>: Low?
    </li>
    <br>
```

```

<li>
    <span class="b">A flawed aid to police.</span><br>
    Having your genome fully available would open you to a small risk of being <a href="#">arrested</a>. This is because your genome could be used to identify you as a criminal or civil disobedient. Anyone living in an authoritarian country or controlled by a totalitarian government would be at risk of being arrested if their genome was found to be that of a criminal or civil disobedient. This is because your genome could be used to identify you as a criminal or civil disobedient. Anyone living in an authoritarian country or controlled by a totalitarian government would be at risk of being arrested if their genome was found to be that of a criminal or civil disobedient.
</li>
<br>
<li>
    <span class="b">DNA framing.</span><br>
    Given a genome, it is not impossible for someone to grow tissue and like plant it in your body. This would be considered a crime of framing. Anyone who pisses off rich criminals, spies, police, or other powerful people would be at risk of being framed. This is because your genome could be used to identify you as a criminal or civil disobedient. Anyone living in an authoritarian country or controlled by a totalitarian government would be at risk of being arrested if their genome was found to be that of a criminal or civil disobedient.
</li>
<br>
<li><span class="b">Releases info about my family members.</span>
    <br>
    The same principle that makes paternity tests so good means that when I disclose my genome, I am also disclosing information about my family members. Any of my antecedents or descendants who fall into any of the categories above would be at risk of being arrested if their genome was found to be that of a criminal or civil disobedient.
</li>
<br>
<li>
    <span class="b">People knowing things about me is creepy!</span><br>
    Please don't look into <a href="{{broker}}>Acxiom or Experian</a>; it will only reveal your genome. Anyone who is obsessed with anonymity would be at risk of having their genome revealed if they were to use these services.
</li>
<br>
</ul>
</div>

```

## NEAR-FUTURE RISKS

```

<ul>
    <li>
        <span class="b">Probabilistic homophobia</span>. <br>
        Homosexuality has a <a href="{{lg}}>strong genetic</a> component; this regularity means that there is a high probability that a large chunk of the world's population is gay. In <a href="{{chunk}}>a large chunk</a> of the world, this is still not information that can be used to identify individuals. However, if this information is combined with other information, such as location and social media activity, it may be possible to identify individuals. This is because your genome could be used to identify you as a criminal or civil disobedient. Anyone living in an authoritarian country or controlled by a totalitarian government would be at risk of being arrested if their genome was found to be that of a criminal or civil disobedient.
        <br><br>
        (I was going to put this risk in the 'actual threats' bit, but the <a href="{{credibility}}>credibility</a> section is more appropriate.)
        <br>
        <i>Risk applies to</i>: Closeted gay people, and straight people who happen to score highly on the homophobia scale.
        <i>Risk to me</i>: Could be.
    </li>
    <br><br>
    <li>
        <span class="b">Mate choice</span>
        <br>
        Once a majority of people are sequenced, we could easily see customs about the genetic makeup of potential mates. This could lead to discrimination based on genetics, such as preferring certain ethnicities or religious groups. Anyone who is discriminated against based on their genetics would be at risk of being treated unfairly.
    </li>

```

(You might think that postwar Europeans are completely inoculated against such things.  
<br><br>(You might think that no one *you* loved would ditch you over some bad genes. A  
<br><br>*Risk applies to*: Anyone who wants kids, or wants to be with someone who does.  
</li>

</ul>

## FAR-FUTURE RISKS

<ul>

<li><span class="b">Clones.</span><br><i>Controversy</i>: People really seem to hate clones for no good reason. (It threatens  
<i>Harm to me</i>: None; more of me is a *good* thing. I am a fragile combination of

I guess they could torture the clone to get to me? But this is silly and I do not negotiate.  
<i>Probability</i>: Of me being cloned off some ancient databank? Minuscule.  
</li>

<br><br>

<li><span class="b">Tailored bioweapons.</span><br>

You might worry about being assassinated with a virus which only kills you. Dunno why that's

<i>Harm to me</i>: No. I am unlikely to sufficiently piss off anyone with a spare billion lives.  
</li>

</ul>

(Then, of course, there's everything I haven't thought of - and everything no-one has thought of - everything which the great random-situation generator called physics hasn't yet thrown up. You should probably seek privacy based on this, more than the above.)

{% include genes/foots.html %}

# Age of Em as graph

Gavin

2017-04-11

1. List claims and implications in AoE
2. Investigate each and every one.
3. Hansonia: Age of Em as visual digraph
4. Turn off / on nodes, colour outcomes.
5. First Order Future: Generalisation where you can play with Bostromian scenario
6. Probabilism

It's worth reading even if you don't care a single jot about the future, since he touches on a sizeable portion of all good settled social science.

# Should we abolish foreign aid?

Gavin

2012-05-01

{% include aid/links.md %}

<br><small>'*The Good Samaritan, after Delacroix*' (1890) by van Gogh</small>

There are eight levels of charity, each greater than the last. The greatest, above which there is no other, is to strengthen another by giving him a gift or free loan... to strengthen his hand until he need no longer beg...

- Maimonides 158

or

... as experience has proved, I believe, without a single exception, that poverty and misery have always increased in proportion to the quantity of indiscriminate charity, are we not bound to infer... that such a mode of distribution is not the proper office of benevolence?

- Thomas Malthus 159

This is my undergraduate economics dissertation, on certain pessimistic analyses of foreign aid. Rather than tackle the impossible question “To what extent has aid caused net socioeconomic development, in general, over 70 years?”, it picks an easy one: “Is aid so bad we need to stop it?”

It has some good bits - I anticipate the “Important/Neglected/Tractable” model, carefully read a bunch of ideologically opposed people with an enormously weird idea, and I note the functional similarities of left-wing and right-wing radicalism. But it’s light on theory; this is history by someone who doesn’t do historiography, macro by someone who couldn’t solve DSGEs, and econometric meta-analysis by a near-illiterate. Even so, I think it shows what literary kinds of research can and can’t do (roughly: great questions, no great answers).

<h3>t1;dr</h3>

<div>

Government-to-government aid was pretty bad for like fifty straight years, 1950 through

Various clever people have claimed aid is so bad it has to stop. (For literally hundreds

```

<ol>
    <li>Aid increased greatly over the C20th.</li>
    <li>Poverty among recipients mostly didn't decrease.</li>
    <li>Nudge nudge wink wink, quite a correlation eh?</li>
</ol>

.) However, the evidence about aid is even weaker than usual for economics, so this claim

Health interventions are the only aid we can confidently say is <i>really</i> good, though

ODA probably is still pretty bad, but people started making encouraging noises in 2005.
</div>

<h3>Introductions</h3>
<div>
    {%
        include aid/intros.html
    %}
</div>

<h3>Why aid?</h3>
<div>
    {%
        include aid/rationale.html
    %}
</div>

<h3>Types of aid abolitionist</h3>
<div>

    <div class="accordion">
        <h3>Peter T Bauer: empiricist, radical</h3>
        <div>
            {%
                include aid/bauer.html
            %}
        </div>

        <h3>William Easterly: empiricist, moderate</h3>
        <div>
            {%
                include aid/easterly.html
            %}
        </div>

        <h3>Dambisa Moyo: ideologue, neoliberal</h3>
        <div>
            {%
                include aid/moyo.html
            %}
        </div>

        <h3>Teresa Hayter: ideologue, neo-Marxist</h3>
        <div>
            {%
                include aid/hayter.html
            %}
        </div>
    </div>
</div>

```

```
        </div>
    </div>

</div>

<h3>Seeking a zero lower bound on aid impact</h3>
<div>
    {%
        include aid/analysis.html
    %}
</div>

<h3>Bibliography</h3>
<div>
    {%
        include aid/biblio.html
    %}
</div>

{%
    include comments.html
}
{%
    include aid/foots.html
}
```

# Ada Lovelace, StackOverflow user #1

Gavin

2017-05-07

Ada, Countess Lovelace, is often called the first programmer. But of the five programs in her notes on the Analytical Engine, four are copies of prior implementations by Babbage.

(Indeed, when we inspect the claim, it is hard to see how Babbage could have designed a processor without mentally coding for it. You can't just throw a bunch of pistons and wheels together and expect it to compute functions.)

This post conducts a static analysis on Analytical Engine pseudocode, to see if we can deem Lovelace the second programmer, or the first StackOverflow user.

(Implement in rotor assembly? Same accumulator, same operation sequence.)

# work space: a calculus of bullshit jobs

Gavin

2017-05-07

What fraction of jobs are bad jobs? What fraction of candidates are bad candidates?

# Regression to the kind

Gavin

2017-05-01

The deep way to study stats is to start off learning a model or a distribution or an estimator and see how they dwindle under a more general method, when you relax its assumptions one by one.

# My idiolect

Gavin

2017-05-31

I have about a million words of my own writing. This is more than enough to infer things about my particular version of my language (my “idiolect”). This is possible because the absolute geniuses at spaCy give away a military-grade English model for free.

- Old blog: 454,000 words (including quotes, drafts, stop words)
- New blog: 20,000 words (including quotes, drafts, stop words)
- Facebook: 200,000 words
- Emails: 10,000 words
- Uni essays: 200,000 words
- My Edinburgh theatre reviews: 2000 words

<http://www.nltk.org/book/> <http://vh216602.truman.edu/agarvey/cs480/nlpcode/mytext.py>

- Median sentence length. Over time.
- Unigram frequency: nltk.FreqDist
- Bigram frequency:
- Rare pairs:
- Unique trigrams:
- Flesch-Kincaid: \* The FK scale is not well-regarded in the field, since it overweights mere vocabulary over clauses and real clarity, but it still gives you a rough picture.
- Topic modelling
- Entity recognition:
  - People
  - Organisations
- Writer invariant
- Then factor out quotations and re-do

CountVectorizer (bag of words model) and TfidfVectorizer (tf-idf weighting for the bag of words model)

{% include idiolect/foots.html %}

# Einstein

Gavin

2018-06-27

Physics becomes in those years the greatest collective work of art of the twentieth century.

Jacob Bronowski

```
{% assign brown = "https://en.wikipedia.org/wiki/%C3%9Cber_die_von_der_molekularkinetischen_Theorie_o
%} {% assign muck = "https://books.google.co.uk/books?id=MjVgeT7Laf8C&pg=PA129&dq=To+think+I+haw
%}
```

*A review of ‘Einstein’ (2007) by Walter Isaacson.*

What to say about the stereotypically great? Start by scrubbing off the accumulated century of journalism and appropriations.

```
<h3>Einstein's scientific achievements</h3>
<div>
  <ul><br>
```

A model of Brownian motion: the decisive argument for the existence of atoms. His model enabled experimental confirmation of Dalton’s theory, after a hundred years of denial or instrumentalism.

An elementary particle, the photon. The atomic hypothesis applied even to light.

A law for the photoelectric effect, implying a quantum theory of all EM radiation. (A realist about quanta, unlike Planck.)

So also lots of pieces of the “old” quantum theory.

A theory of light and so space and time, special relativity.

A physical constraint on metaphysics: no absolute time.

A fairly consequential law, mass-energy equivalence

A flawed but progressive theory of heat capacity, the Einstein theory of solids

A better method of analysing quantum systems, “EBK”. An ignored semiclassical precursor to quantum chaos theory.

The greatest scientific theory, General Relativity. Explaining gravity and, so, the shape of the universe.

Implies the first modern cosmology

Gravitational lensing (confirmed 1998)

Inadvertently predicted dark energy.

A crucial experiment: gravitational waves. (Confirmed 2015.)

Implies a whole lot more like black holes but you can't name everything "Einstein thing".

A general method for thermodynamics and information theory: Bose-Einstein statistics.

New state of matter: the Bose-Einstein condensate

Fruitful failed theory: first local hidden variable theory

A profound phenomenon, quantum entanglement. (Susskind calls entanglement "Einstein's last great discovery", though he 'discovered' it by trying to reduce away Copenhagen interpretation, taking entanglement to be a disproof.) (Confirmed properly 2015.)

A crucial experiment for a metaphysical principle, local realism is false!: EPR

Inadvertently, a physical constraint on metaphysics: nonlocality.

Thought-experiment: The content of the "Schrödinger's" cat setup

Repostulation of wormholes. (Not confirmed.)

Isotope separation methods for the Manhattan project.

Also a nontoxic fridge

Besides his own prize, confirmations of Einstein's theories have led to 4 Nobel Prizes (1922, 1923, 1997, 2001) so far, and first-order extensions several more (1927, 1929, 1933, 2020 at very least). We should expect a few more, for grav waves and not inconceivably for wormholes, some day.

Isaacson, like most people, portrays Einstein's post-1935 work as a dogmatic waste - he spent about thirty years straining to produce a field theory that could get rid of the spookiness and probabilism of QM. If you compare the output of the first half of his life to the second, sure it looks bad. But he was giving classical physics (determinism, continuousness, simplicity, fierce parsimony, beauty-based reasoning) a well-deserved last shake.

Imagine maintaining full-time effort over thirty years of failure, with your whipsmart peers all tutting and ignoring you. His unified field efforts are methodologically sort of like string theory: a hubristic search over mathematical forms without contact with the actually physical to help limit the formal space.

He had a decent decision-theoretic argument for his doomed crusade:

When a colleague asked him one day why he was spending — perhaps squandering — his time in this lonely endeavor, he replied that even if the chance of finding a unified theory was small, the attempt was worthy. He had already made his name, he noted. His position was secure, and he could afford to take the risk and expend the time. A younger theorist, however, could not take such a risk, for he might thus sacrifice a promising career. So, Einstein said, it was his duty to do it.

People also try to attach shame to him for his wildly stubborn anti-Copenhagen crusade: years spent thinking up tricky counterexamples for the young mechanicians, like an angry philosopher. But I think he had a good effect on the discourse, constantly calling them to order, and leaving it clear, after all, that it is a consistent view of the evidence.

The only unforgiveable bit in his later conservatism is that he ignored the other half of the fundamental forces, the strong and weak forces, and for decades. Two forces was hard enough to unify. I suppose another point against his long, long Advanced Studies is that he could have done even more if he had helped push QM along; as late as 1946, Wheeler tried to convince him to join in. As it is we have evidence against the unified field: “Einstein failed”.

What was so moral about him? Well, he was ahead of his time (still is):

Denounced WWI as the senseless crap it was.

Never went to the Soviet empire (despite repeated invites).

Denounced the Nazis from '31, despite/because of public threats to his life.

Flipped from pacifism at the right moment.

Many early actions for US civil rights, including work against McCarthyism.

Sold his original manuscripts for War Bonds

Even his Zionism was enlightened (pro-migration, anti-state, anti-Begin):

“Should we be unable to find a way to honest cooperation and honest pacts with the Arabs,” he wrote [Chaim] Weizmann in 1929, “then we have learned absolutely nothing during our 2,000 years of suffering.”

He proposed, both to Weizmann and in an open letter to an Arab leader, that a “privy council” of four Jews and four Arabs, all independent-minded, be set up to resolve any disputes. “The two great Semitic peoples,” he said, “have a great common future.” If the Jews did not assure that both sides lived in harmony, he warned friends in the Zionist movement, the struggle would haunt them in decades to come. Once again, he was labeled naïve.

Einstein is like Bertrand Russell, only much more so: even more brilliant, even more rebellious, even more politically active, even more aloof, even more relentless, even more neglectful of his family. (Russell on hearing relativity for the first time: “To think I have spent my life on absolute muck.”)

Along with Ibn Rushd, Pascal, Leibniz, Darwin, Peirce, Russell, Turing, Chomsky, Einstein is one of our rare complete intellectuals: huge achievements in science, beautiful writing, good jokes, original philosophy, moral seriousness. To have warmth too, as Einstein does abundantly, doesn't have much precedent. However much Einstein is misattributed vaguely pleasant, vaguely droll, vaguely radical statements, the fact is he actually was pleasant, funny, radical. Believe the hype.

The usual word for this lot is 'polymath' - but though we are mad keen on polymaths, their generalism is seen as a laudable extra, rather than the vital service I now think they alone can give: you want people who have proven they can discover things to tackle your ancient ill-defined questions (beauty, goodness, justice, existence). The above are more than subject-matter polymaths; they are both thinkers and doers, hackers and painters, servants and masters, above their time and ahead of it.

You can't do good unless you know a great deal about the targets of your morals; you want the vast imaginative search over philosophical possibilities to be aided by what we actually know. (As the noted writer against scientism, Ludwig Wittgenstein put it:

Is scientific progress useful for philosophy? Certainly. The realities that are discovered lighten the philosopher's task: imagining possibilities.

)

<h3>Other greats</h3>

<div>

Maxwell, Boltzmann, Schrödinger, and Feynman basically fit the above: they are as good as Goethe tried admirably, but didn't achieve much science. Descartes should be on there but

</div>

One particularly charming bit in this book covers Einstein's long friendship with the Queen Mother of Belgium. When Szilard warns him that nuclear fission has been achieved and could give the Nazis dominion over all, Einstein's first thought is to ask Elisabeth to sort it out, by grabbing all the Central African uranium and sending it far from the Nazis. (As it happens, the Uranverein got their uranium from Czechoslovakia.)

Isaacson read all the letters, formed a view on all the academic controversies (Maric's contribution, baby Lieserl, what sort of deist or Zionist or pacifist he was), and covers most of the papers, recasting the classic thought experiments very lucidly. This was a huge pleasure. Read with Wikipedia open, though: C20th physics is way too deep and broad for one book.

<h3>Why listen to me on this topic?</h3>

<div>

<i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect

<ol>

<li>immersion in the field and/or good priors for what makes for an extraordinary story<br/><li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes

<li>incredible amounts of argument-checking, which doesn't need domain knowledge  
</ol>  
I always try to do (3) but surely often fail.</i> <br><br><br>

In this case, don't trust me much. I am no physicist, and only half a scientist. I look  
</div>

# How lethal are the Tories? Part 1

Gavin

2018-02-25

{% include killer-tories/links.md %}

You sometimes see the claim that the Conservative-led coalition killed thousands of disabled people by spuriously cutting them off from disability benefit.<sup>3</sup>

This is naive to the point of deceit, as I'll show. But it's based on something which sounds similar: the fact that between 2011 and 2014, 2,380 people died after being declared "fit for work" (FFW), i.e. after having their main income stopped. (If we were to establish causation, and so responsibility, this would make the Tories about as lethal as uterine cancer, at 720 UK deaths a year.)

It's not hard to find cases where causation seems likely. But, by inferring causation from the above raw figure, the "2,380" claim implies that the Conservatives are responsible for all mortality during their reign - which, even speaking as a Scotsman, seems a bit strong.

Media discussion of this fact was sloppy even by the low standards of public policy discussion.<sup>11</sup> There is no justice without accuracy.

*Terms:*

- ESA : Employment and Support Allowance; the UK's newish main disability benefit.
- WCA : Work Capacity Assessment. Quasi-medical screening process for ESA. Introduced by Labour in 2008, made much stricter by the Conservatives in 2011.

The following is just an observational argument: it doesn't exonerate or condemn. All I can say for it is that it's less pig-ignorant than parroting the uncontrolled figure. If you take one thing from this, make it *You cannot infer anything about impact from one number, at one point in time, without a reference class.*

In particular, it doesn't make sense without accounting for the number of deaths in this group *before* the WCA reform. (Maybe 2380 is an improvement.) And it doesn't make sense to compare even those numbers without accounting for large known influences on mortality, e.g. seeing if ages and genders differ between the compared groups. What we actually need is not 'deaths' but 'excess' deaths.

It took me 10 seconds to find age-adjusted data, compared to the general population, before and after WCA:

```

```

```
<small>Suggested headline: <i>"Go on the dole to save your life!"</i> <a href="#fn:2" id="fnref:2">2</a>
```

No large changes: people on incapacity benefit have been dying very slightly less (1043 -> 1032), and there's a slight increase (116 to 138) among JSA recipients. Given ~2.5m people on ESA, this fall works out to about 275 fewer deaths per year. 5 4

But we're interested in the ones who aren't on disability any more; in particular, the ones who were kicked off. (Many move onto Jobseeker's Allowance (JSA), which is actually the lowest-mortality group, even after adjusting for the relative youth of people on JSA. Then there's a group who presumably fall off the official stats entirely.)

So, compare the mortality rate of people on ESA (1.032%) with those kicked off it. The published data for WCA results only goes up to March 2013 at present; I'll update this when they're out, but for now let's plot a dumb model for the 2013-4 rate:

```

<a href="#fn:6" id="fnref:6">6</a>
```

- *December 2011 to March 2013*: 238,100 declared fit for work.
- *Extrapolation for April 2013 to February 2014*: 131,500, if trend continued.
- *Estimated total “fit for work”, December 2011 to February 2014*: 369,600.
- *December 2011 to February 2014*: 2,380 deaths among “fit for work” within 6 months of decision.
- *Non-age-standardised death rate among “fit for work”*: 0.64%.
- *Age-standardised death rate among ESA recipients*: 1.03%.

So the *non-age-standardised* death rate among those declared fit to work (0.64%) was halfway between the unfit-to-work (1.03%) and the general population rate (0.24%). 10

```

```

What does this tell us? That the “fit-for-work” population is not the same as the general population in some way. Without age standardisation, the following explanations are equally consistent:

- People deemed “fit-for-work” were generally older than the general population.
- The “fit-for-work” consist of more men than the general population does.

- Half of “FFW” people were as unwell as the “unfit-for-work”; all “FFW” people were half as healthy as the general population; more likely, some mixture of these health statuses. This would be an indictment of WCA, since the general population is exactly what they’re treated as being.
- Some combination of the above.

If the “FFW” had the same age and health distribution as the general population, you’d expect them to suffer roughly 887 deaths a year.<sup>7</sup> As it is, there were 1057, or something like ~170 excess deaths a year.<sup>8</sup>

What we can get from this is an *upper bound* on responsibility. If we insist on extracting a figure to compare to the reported figure, then the WCA is associated with at most 383 deaths over this period, and probably less.<sup>9</sup> “2,380” is many times too high, *even if* it had been stated as an honest observation and not the resounding proof of blame it was stated as.

(Clearly this is too ambiguous for the purposes of political point-scoring. Things often are.)

The above has nothing to say about causation; many other things besides WCA could have and will have borne on these (e.g. age distribution, accidents, violence, decompensation). I don’t even have row-level data to properly establish that FFWs are a different population, let alone enough to isolate WCA’s effects on them. Actually all you’d need is the sample variances, but I can’t see them.

---

## Labour vs Tories

Some people have told me that the above is incomplete without a comparison to the Labour period (2008-11); that it reads like a shrug. That’s wrong - it is reasonable and helpful to remove individual falsehoods from the pool - but neither is it unfair.

There were 537,800 “fit-for-work” judgments between October 2008 and January 2011.

We don’t have death data for them - the relevant Freedom of Information report for the period gives us the following shrug:

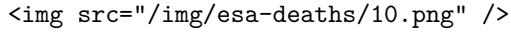
Data on the number of incapacity benefits claimants that have died following a fit for work decision is not available, as the Department does not hold information on a death if the person has already left benefit.

That’s not a lie, but we know it to be half-arsed given that they came up with this data for a different window, two years later.

Here’s something even more circumstantial than my above analysis:

```

```



Too much going on there; we just don't know.

---

## How bad are/were work competency assessments?

*This section has several made up numbers.*

The main reasons to be suspicious of the 2011 WCA are: 1) they are sometimes not conducted by medical staff; 2) the private companies that run them are given narrow norms that probably result in a de facto quota; 3) they penalise less visible conditions like major depression and chronic pain.

If we had just a couple of numbers, we could use the awesome machinery of the confusion matrix to objectively rate how good WCAs are at their allotted dirty job.

Buckle up, because it's time for some Bayesian inference.

If the WCA is a disability test, then call a fit-for-work judgment a ‘negative’ result: i.e. the WCA test doesn’t think you are disabled enough. Assume that a successful appeal is the same as showing a false negative on the original test (though in fact appeals will have some error rate too).

- Base rate for disability  $P(H)$ : One estimate is 21% of UK adults.
- False positive  $P(E \mid \sim H)$ : being flagged unfit-for-work despite not being disabled. Probably low: 10% ?
- False negative  $P(\sim E \mid H)$ : being flagged fit-for-work despite being disabled: FNR = 59%
- True positive  $P(E \mid H)$ : being flagged unfit-for-work and being disabled:  $1 - \text{FNR} = 41\%$
- True negative  $P(\sim E \mid \sim H)$ : being flagged fit-for-work and not being disabled:  $1 - \text{FPR} = 0.9$

We can use these to guess the conditional probability that someone is disabled given a positive WCA result (“unfit-for-work”):

$$\begin{aligned} P(H \mid E) &= P(E \mid H) \times P(H) / P(E \mid H) \times P(H) + P(E \mid \sim H) \times (1 - P(H)) \\ &= (0.41 \times 0.21) / (0.41 \times 0.21 + 0.1 \times 0.79) \\ &= 52.2\% \end{aligned}$$

Slightly better than a coin flip; and the conditional probability that someone is disabled in spite of a negative WCA result (“fit-for-work”):

$$\begin{aligned} P(H \mid \sim E) &= P(\sim E \mid H) \times P(H) / P(\sim E \mid H) \times P(H) + P(\sim E \mid \sim H) \times P(\sim H) \\ &= (0.59 \times 0.21) / (0.59 \times 0.21 + 0.9 \times 0.79) \\ &= 16\% \end{aligned}$$

i.e. Under these estimates, the test is fairly weak evidence. (Don’t rely on this; there are too many assumptions, and of necessity I’ve used the UK population

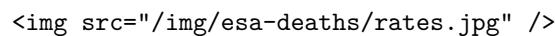
rather than the test-taking population, which is bound to have a higher base rate.)

---

## Data

People struggled with the Department for Work and Pensions to get these figures published. This is sometimes read as an admission of guilt. But given how naively the 2,380 figure was received, it is hard to blame them for their cowardice.

Although this official figure is very misleading



since it hides a recent rise in death rate (2010 - 2013) behind the big drop between 2003 - 2008.

We could reduce our uncertainty if we had data on the cause of death - e.g. hypothermia and suicide being evidence of WCA responsibility, while decompensation of chronic illnesses wouldn't be. But we don't.

I had a look for people who disappear from the system entirely, among the homeless. The government doesn't collect this information (an oversight I'm inclined to be cynical about), and the charity Crisis haven't updated their numbers since 2009, just outside our analytical window. They are not counted; they do not count.

This analysis doesn't cover public time lost to bureaucracy, nor the poverty and distress of those who didn't die.

---

Part 2 of this series will be about the stronger academic claim that austerity caused 30,000 - 45,000 excess deaths, mostly among the elderly.

Part 3 might be about the mental health impact of WCA, though maybe not - that report doesn't make very questionable claims.

{% include killer-tories/foots.html %}

# Three Tsars

Gavin

2017-05-07

(War toll, domestic murders, extent of secret police, foreign espionage, theft, growth)

# ‘Hitler’s Uranium Club’ (2008) by Bernstein

Gavin

2017-06-01

There are few, if any, other instances in recorded history where we have the conversations of leading figures as they complete one era, come to terms with it, and prepare their strategy for the next. It is as though these men were lifted out of history at a crucial turning point—from the age of conventional weapons to the nuclear era — placed within a timeless container and told to discuss their past and future as the recorders roll.

- Jeremy Bernstein

Astonishingly dramatic; also as pure as primary sources get. Months of secret eavesdropping on the German nuclear scientists, including just as they hear of Hiroshima and so their eclipse. Innocent of the microphones, the men concede their ignorance without ego, their character without any obfuscating propriety.

There are still two impurities: their words are both transcribed and translated by strangers (the German originals were destroyed). The physicists speak to us here in full sentences, with little of the fragmentariness and repetition of real speech. And it takes someone as highly trained as Bernstein to get us over the technical barrier. Even so, this is as plain and self-interpreting as history gets. For six months these men play madlibs, argue, and run around the garden, while the English and we listen in.

Two of them are unjustly detained: Hahn is a sweetheart and von Laue a droopy hero. The Party functionary Diebner is a funny guy, even though he has the most responsibility for the Nazi weapons project. Harteck is the most technically astute by far: he guesses a huge amount correctly, all in the teeth of loud ignorance by his more prestigious peers. von Weizsacker is the slimmest. Heisenberg is just weird: there's a faint echo of the clear-sight-and-moral-vacuum of Eichmann. Enormous intelligence and no sense.

The morality of their wartime actions does not come up very much (except when raised by sweetheart Hahn or von Laue). They are mostly glad of the destruction of the Nazis, and Wirtz is horrified by the scale and singularity of SS murder. But the rest are more self-regarding than pro or anti Nazi. (Again, it is wonderful to read these and actually know they meant it.)

(What about the morality of our reading the reports? I don't have a clear opinion, but doing so after their deaths seems mostly fair.)

They very often speak about money, Heisenberg in particular. (Not just research funding or aid for their families in Occupied Germany, but dolla dolla bills.) On hearing that Hahn had won a Nobel:

"it says that you are supposed to receive the Nobel Prize for 1944." The excitement that struck the ten detainees at this moment is hard to describe in a few words. Hahn did not believe it at first. In the beginning he turned away all the offers of congratulations. But gradually we broke through, with Heisenberg in the lead, who congratulated him heartily on the 6200 pounds.

Bernstein's editorial voice is a bit strong. Representative footnotes on the transcripts:

But his other qualities are huge and unique: he knew some of the protagonists personally, and worked on nuclear weaponry himself. He is out to get Heisenberg, and overreads a few times. But this is because people (Powers, Frayn to a degree) persist in rose-tinting him: there's this idea that Heisenberg feigned incompetence at reactor-making as anti-Nazi activism. The transcripts make clear that he'd have made a bomb if he could, not because he is a Nazi or a German but because he was amorally curious, and hungry for primacy. Heisenberg does object to Nazism. But not very strongly.

Bernstein's conclusion is that the project was pretty much a shambles. They had a two-year head start on the Allies, but failed for several reasons: they had < 1% of the funding of the Manhattan Project, an unbelievably bad administration and communication of data and ideas, and key resources like deuterium kept getting bombed. But Bernstein feels able to go for the jugular:

reading this lecture, I am once again struck by the intellectual thinness of this group. Here are ten German nuclear scientists — nine if one does not count von Laue — who are supposed to be the cream of the crop, the intellectual elite, of German nuclear physics, men who had been working on these questions for several years. And look at the discussion it produced.

To see what I have in mind, let us entertain the following fantasy. Suppose the tables had been turned and ten of the best Allied scientists had been interned in Göttingen when a hypothetical German atomic bomb went off. Whom shall we include? Fermi, Bethe, Feynman, Serber, Wigner, von Neumann, Oppenheimer, Peierls, Ulam, Teller, Bohr, Frisch, Weisskopf... What would the technical conversation have been like? No doubt there would have been disagreements and some fumbling. But like this? The question answers itself.

Yet even with these handicaps, it looks like Harteck could have built a basic pile in 1940, if the project was headed by someone less arrogant than Heisenberg. And that pile would have brought all the funding, and maybe sorted out their many collective muddles and lack of engineering care.

5/5 for Bernstein's commentary and the hair-raising fact of their existence.

<h3>Why listen to me on this topic?</h3>

<div>

<i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect

<ol>

<li>immersion in the field and/or good priors for what makes for an extraordinar

<li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes

<li>incredible amounts of argument-checking, which doesn't need domain knowledge

</ol>

I always try to do (3) but surely often fail.</i> <br><br><br>

In this case, don't trust me much. I am no physicist, and only half a scientist.

</div>

The 2009 PhilPapers Survey

A priori knowledge: yes or no?

Accept: yes (71.1%)

Lean toward: yes

Accept: no (18.4%)

Lean toward: no

Other (10.5%)

Yes: There exist facts we can know without our knowledge being based on sensory experience. No: Justification of knowledge requires sensory experience.

Abstract objects: nominalism or Platonism?

Accept: nominalism (37.7%)

Lean toward: nominalism

Accept: Platonism (39.3%)

Lean toward: Platonism

Other (23.0%)

Abstract objects are objects independent of any physical pattern in space-time.  
Nominalism: Abstract objects do not exist. Platonism: Abstract objects exist.

Aesthetic value: objective or subjective?

Accept: objective (41.0%)

Lean toward: objective

Accept: subjective (34.5%)

Lean toward: subjective

Other (24.5%)

Analytic-synthetic distinction: yes or no?

Accept: yes (64.9%)

Lean toward: yes

Accept: no (27.1%)

Lean toward: no

Other (8.1%)

Yes: Some sentences are true solely due to the meanings of the words. No: Every sentence is open to empirical falsification or no sentence is open to falsification.

Epistemic justification: internalism or externalism?

Accept: internalism (26.4%)

Lean toward: internalism

Accept: externalism (42.7%)

Lean toward: externalism

Other (30.8%)

Externalism: Belief can be justified even when the justification is not consciously available to the subject. Internalism: Belief is only justified if there is conscious understanding of the justification.

External world: idealism, skepticism, or non-skeptical realism?

Accept: idealism (4.3%)  
Lean toward: idealism  
Accept: skepticism (4.8%)  
Lean toward: skepticism  
Accept: non-skeptical realism (81.6%)  
Lean toward: non-skeptical realism  
Other (9.2%)

Idealism: Reality is not mind-independent. Skepticism: Mind-independent reality exists, but we lack epistemic access to it. Non-skeptical realism: Mind-independent reality exists, and we have epistemic access to its structure.

Free will: compatibilism, libertarianism, or no free will?

Accept: compatibilism (59.1%)  
Lean toward: compatibilism  
Accept: libertarianism (13.7%)  
Lean toward: libertarianism  
Accept: no free will (12.2%)  
Lean toward: no free will  
Other (14.9%)

Compatibilism: We can have free will in a deterministic universe. Libertarianism: Incompatibilism is true and we have free will. No free will: Free will does not exist.

God: theism or atheism?

Accept: theism (14.6%)  
Lean toward: theism  
Accept: atheism (72.8%)  
Lean toward: atheism  
Other (12.6%)

Theism: Gods exist. Atheism: Gods do not exist.

Knowledge: empiricism or rationalism?

Accept: empiricism (35.0%)  
Lean toward: empiricism  
Accept: rationalism (27.8%)  
Lean toward: rationalism  
Other (37.2%)

Empiricism: Only sensory experience gives us new information. Rationalism: Some information exists that we can arrive at without sensory experience.

Knowledge claims: contextualism, relativism, or invariantism?

Accept: contextualism (40.1%)  
Lean toward: contextualism  
Accept: relativism (2.9%)  
Lean toward: relativism  
Accept: invariantism (31.1%)  
Lean toward: invariantism  
Other (25.9%)

Contextualism: The truth of a knowledge claim depends on the context in which it is uttered. Relativism: Whether a subject possesses knowledge of a proposition is relative to a set of epistemic standards. Invariantism: The truth of knowledge claims does not depend on context and is not relativized to epistemic standards.

Laws of nature: Humeanism or non-Humeanism?

Accept: Humeanism (24.7%)  
Lean toward: Humeanism  
Accept: non-Humeanism (57.1%)  
Lean toward: non-Humeanism  
Other (18.2%)

Humeanism: The laws of nature are compressed descriptions of salient patterns in the distribution of physical events. Non-Humeanism: The laws of nature are not mere descriptions, but actually determine the distribution of physical events.

Logic: classical or non-classical?

Accept: classical (51.6%)  
Lean toward: classical  
Accept: non-classical (15.4%)  
Lean toward: non-classical  
Other (33.1%)

Classical: Standard logics, such as Boolean logic or first-order predicate calculus, are best (or correct). Non-classical: The best logic is not classical (e.g., paraconsistent logic).

Mental content: externalism or internalism?

Accept: externalism (51.1%)  
Lean toward: externalism  
Accept: internalism (20.0%)  
Lean toward: internalism  
Other (28.9%)

Externalism: The representational content of our mental states is dependent upon properties of our external environment. Internalism: The representational content of our mental states is fixed by our brain state.

Meta-ethics: moral realism or moral anti-realism?

Accept: moral realism (56.4%)  
Lean toward: moral realism  
Accept: moral anti-realism (27.7%)  
Lean toward: moral anti-realism  
Other (15.9%)

Moral realism: Objective moral facts exist. Moral anti-realism: Objective moral facts do not exist.

Metaphilosophy: naturalism or non-naturalism?

Accept: naturalism (49.8%)  
Lean toward: naturalism  
Accept: non-naturalism (25.9%)  
Lean toward: non-naturalism  
Other (24.3%)

Naturalism: All causes are natural. Non-naturalism: Supernatural causes exist.

Mind: non-physicalism or physicalism?

Accept: non-physicalism (27.1%)  
Lean toward: non-physicalism  
Accept: physicalism (56.5%)  
Lean toward: physicalism  
Other (16.4%)

Physicalism: A physical duplicate of our world must necessarily also be a mental duplicate. Non-physicalism: Mental states are not dependent on physical states.

Moral judgment: cognitivism or non-cognitivism?

Accept: cognitivism (65.7%)  
Lean toward: cognitivism  
Accept: non-cognitivism (17.0%)  
Lean toward: non-cognitivism  
Other (17.3%)

Cognitivism: Moral statements have truth conditions. Non-cognitivism: Moral statements have no truth conditions.

Moral motivation: internalism or externalism?

Accept: internalism (34.9%)  
Lean toward: internalism  
Accept: externalism (29.8%)  
Lean toward: externalism  
Other (35.3%)

Internalism: A necessary connection exists between sincere moral judgment and either justifying reasons or motives. Externalism: Any connection that exists between moral judgment and motivation is purely contingent.

Newcomb's problem: two boxes or one box?

Accept: two boxes (31.4%)  
Lean toward: two boxes  
Accept: one box (21.3%)  
Lean toward: one box  
Other (47.4%)

Omega appears before you with two boxes and says you may take Box A or take both Box A and Box B. Omega has almost certain predictive power and does not lie. Omega has predicted which you will choose; if Omega predicts you will take just Box A, then Box A will contain \$1,000,000. Box B always contains \$1,000. How many boxes do you take?

Normative ethics: consequentialism, deontology or virtue ethics?

Accept: consequentialism (23.6%)  
Lean toward: consequentialism  
Accept: deontology (25.9%)  
Lean toward: deontology  
Accept: virtue ethics (18.2%)  
Lean toward: virtue ethics  
Other (32.3%)

Consequentialism: The morality of actions depends only on their consequences.  
Deontology: There are moral principles that forbid certain actions and encourage other actions purely based on the nature of the action itself, not on its consequences. Virtue ethics: Ethical theory should not be in the business of evaluating actions, but in the business of evaluating character traits.

Perceptual experience: disjunctivism, qualia theory, representationalism, or sense-datum theory?

Accept: disjunctivism (11.0%)  
Lean toward: disjunctivism  
Accept: qualia theory (12.2%)  
Lean toward: qualia theory  
Accept: representationalism (31.5%)  
Lean toward: representationalism  
Accept: sense-datum theory (3.1%)  
Lean toward: sense-datum theory  
Other (42.2%)

Disjunctivism: In normal cases, when a person is perceiving something, the object of their perception is a mind-independent object. Representationalism: Perceptual experience is representational. Sense-datum theory: The objects of our perception are not mind-independent entities, they are mind-dependent objects called sense-data. Qualia theory: The phenomenal character of our perceptual experience is non-representational.

Personal identity: biological view, psychological view, or further-fact view?

Accept: biological view (16.9%)

Lean toward: biological view

Accept: psychological view (33.6%)

Lean toward: psychological view

Accept: further-fact view (12.2%)

Lean toward: further-fact view

Other (37.3%)

Physical view: The maintenance of personal identity requires bodily continuity.

Psychological view: The maintenance of personal identity requires continuity of psychological states.

Politics: communitarianism, libertarianism, or egalitarianism?

Accept: communitarianism (14.3%)

Lean toward: communitarianism

Accept: libertarianism (9.9%)

Lean toward: libertarianism

Accept: egalitarianism (34.8%)

Lean toward: egalitarianism

Other (41.0%)

Proper names: Fregean or Millian?

Accept: Fregean (28.7%)

Lean toward: Fregean

Accept: Millian (34.5%)

Lean toward: Millian

Other (36.8%)

Fregean: The meaning of a proper name is a way of conceiving of its bearer.

Millian: The meaning of a proper name is its bearer.

Science: scientific anti-realism or scientific realism?

Accept: scientific anti-realism (11.6%)

Lean toward: scientific anti-realism

Accept: scientific realism (75.1%)

Lean toward: scientific realism

Other (13.3%)

Scientific anti-realism: There are no strong reasons to believe in their theoretical claims about unobservable entities (though epistemic justification of predictions exist). Scientific realism: There are strong reasons to believe in the theoretical claims about unobservable entities made by our best scientific theories.

Teletransporter (new matter): survival or death?

Accept: survival (36.2%)

Lean toward: survival

Accept: death (31.1%)

Lean toward: death

Other (32.7%)

You are placed in a machine that will instantaneously disintegrate your body, in the process recording its exact atomic configuration. This information is then beamed to another machine far away, and in that machine new matter is used to construct a body with the same configuration as yours. Would you consider yourself to have survived the process, and teleported from one machine to the other ("survival")? Or do you think you have died, and the duplicate in the far away machine is a different person ("death")?

Time: B-theory or A-theory?

Accept: B-theory (26.3%)

Lean toward: B-theory

Accept: A-theory (15.5%)

Lean toward: A-theory

Other (58.2%)

B-theory: Specifying the temporal ordering of all events in space-time exhausts all the objective temporal facts about those events. A-theory: Specifying the temporal ordering of all events in space-time does not exhaust all the objective temporal facts about them.

Trolley problem (five straight ahead, one on side track, turn requires switching): switch or do't switch?

Accept: switch (68.2%)

Lean toward: switch

Accept: don't switch (7.6%)

Lean toward: don't switch

Other (24.2%)

There is a trolley traveling along a set of tracks. The driver has lost control of the trolley. On the track ahead of the trolley are five people who cannot get off the track in time and will all die if the trolley gets to them. You are standing next to a lever that can switch the track the trolley will take, preventing the deaths of the five people. On the other track is a single person who also cannot get away in time and so will die if you switch the track. Do you refrain from switching the track or do you switch the track?

Truth: correspondence, deflationary, or epistemic?

Accept: correspondence (50.8%)

Lean toward: correspondence

Accept: deflationary (24.8%)

Lean toward: deflationary

Accept: epistemic (6.9%)

Lean toward: epistemic

**Other** (17.5%)

Correspondence: A proposition is true if and only if it bears some sort of congruence relation to a state of affairs that obtains. Deflationary: Ascribing truth to a proposition amounts to no more than asserting the proposition. Epistemic: To say that a proposition is true is just to say that it meets a high standard of epistemic warrant, and that we are thereby justified in asserting it.

Zombies: inconceivable, conceivable but not metaphysically possible, or metaphysically possible?

Accept: inconceivable (16.0%)

Lean toward: inconceivable

Accept: conceivable but not metaphysically possible (35.6%)

Lean toward: conceivable but not metaphysically possible

Accept: metaphysically possible (23.3%)

Lean toward: metaphysically possible

Other (25.1%)

A zombie is physically identical to a human being but does not possess phenomenal experience. There is nothing it is like to be a zombie.

Inconceivable: We cannot fully conceive of a zombie. Conceivable but not metaphysically possible: One can arrive at a coherent conception of zombies, but objects that match this conception cannot possibly exist, not even in worlds with different laws of nature than ours. Metaphysically possible: The existence of zombies is possible.

# Index of Global Deaths, 2016

Gavin

2017-06-25

# Disambiguating the first computer

Gavin

2017-08-28

```
{% include comput/links.html %}

<center><br>
    Specify what you mean by computer, and I'll tell you the first computer: <br>
    <i>(click any radio button to start)</i><br><br>
</center>

{% include comput/app.html %}
```

Let me emphasize that there is no such thing as “first” in any activity associated with human invention. If you add enough adjectives to a description you can always claim your own favorite. For example the ENIAC is often claimed to be the “first electronic, general purpose, large scale, digital computer” and you certainly have to add all those adjectives before you have a correct statement...

– Michael Williams

People think they know what they mean when they say “computer” - it’s the thing with a screen and a mouse that gives you cat photos. In that narrow sense, the SDS 940 (1968) that ran Engelbart’s On-Line System was ‘the first computer’.

This is obviously no good: it disqualifies a hundred years of earlier digital computers. Luckily, the name’s a clue: computers are things that do computations. However, all of reality can be usefully considered as computation. So a computer can’t be just “a system which transforms physical information”, because everything does that.

Data: A randomly selected atom is not a computer. A gun is not a computer. An abacus is not a computer, nor is its descendent the slide rule. A primate doing addition is not the kind of computer we are talking about. So we want the first inorganic device that can do complex information-processing automatically, on demand. 2

(Electricity isn’t key, though. The fact that we use voltages to conduct most of our computations is a matter of convenience, not essence.)

When asking “what was the first computer?”, people usually mean the first modern computer, where “modern” is some collection of the following properties:

fully-electronic, Turing-complete, stored-program, binary-digital, parallelised, integrated-circuit-transistorised, virtual-memory, instruction-set-architecture, presenting a desktop metaphor to the user.

{% include comput/details.html %}

{% include comput/foots.html %}

# Breaking changes in key open-source projects

Gavin

2017-06-25

# The typical language of postmodernism

Gavin

2011-08-15

{% assign phob = “<https://www.jstor.org/stable/24440248>” %}

Where questions of style and exposition are concerned I try to follow a simple maxim: if you can't say it clearly you don't understand it yourself. - John Searle

Sometimes the obvious is the enemy of the true. - Gabriel Stolzenberg

I am about to begin studying Derrida, because I wish to know if he's the most brilliant comedian who has ever lived. But, while reading around him, I find that the first thing noted about him is not his extraordinary reinterpretations of canon philosophy; not his seminal critiques of structuralism and phenomenology; nor even that he was continually demonized as a nihilist: the first thing about him is that he could not (or would not) write well. And, bizarrely, this foible is the most popular dismissal of the unsettling work that Derrida and other theorists in that intellectual direction have produced. Taking John Searle's personal maxim (cited above) as given, the critique seems to run: “they write obscurely, therefore they are all speaking nonsense. I don't follow, therefore it doesn't follow.”

It is suggested that this obscure “they” are a subculture in academic thought, uniting the disciplines Cultural studies, Hermeneutics, Post-colonial studies, Queer theory, Gender studies and Critical theory (or just Theory), but also work of any discipline associated with the buzzwords/methods “postmodernism”, “poststructuralism”, “La pensée 68”, “Lacanian psychoanalysis”, and “deconstructionism”. I hope that this grouping looks as artificial to you as it does to me, but for the sake of this let's lend it subsistence under the umbrella postmodernism.

It is suggested that doctrines held in common in this assortment are now the ruling tendency in some humanities departments, and it is further suggested that the people identified with it have nothing to them; that they're intellectually bankrupt or counter-productive. The backlash has been ongoing for perhaps thirty years, with a great number of clearly-written things in popular philosophy, popular science, Marxist philosophy, and journalism arguing that the whole (putative) thing is an outbreak of disguised scepticism, anti-rationalism, or, at the most declamatory, shoddy pseudo-philosophy.

One might call these critics of postmodernism pomophobes; critiques based in postmodernism's writing style are thus cases of aesthetic, or rhetorical pomophobia. "Postmodernism" defies definition in part because it is massively diverse (and potentially bogus), in part because the term is often used pejoratively and disowned by those so-labelled. More fundamentally, definitions fail because the act of definition is seen to be self-defeating: to define is to set a fixed semantic limit, and a recurring thesis is exactly the denial of such "objective" meaning. I offer common symptoms of postmodernism: Breathtaking openmindedness:

1. Semiotics as key method (much time spent critiquing the language and symbolism used, as well as the actual positions)<sup>34</sup>
2. Semantic relativism (meaning as necessarily contextual, "destabilized")<sup>5</sup>
3. Interpretative pluralism as to texts (we apply, not extract meaning).
4. Cultural relativism (rejection of "imperialism", comparative hierarchy)<sup>6</sup>  
Reality viewed as a product of texts (i.e. inescapably structured by our stories, theories, and values, and thus forming enclosed "metanarratives"):
5. Rejection of epistemic objectivity. (Knowledge as only meanings.)<sup>7</sup>
6. Anti-foundationalism<sup>8</sup>
7. Heavy emphasis on socialization and ideology as determiners of identity and meaning. Also the rejection thereof, as in Lyotard's: "I define postmodern as incredulity toward metanarratives."<sup>9</sup>
8. Belief in the social importance of Theorists.
9. Belief in critical reading as a political act.<sup>10</sup> Political radicalism:
10. Anti-scientism (from 4 and 5)
11. Anti-humanism (individual conceived as mere bundle of socializations)<sup>11</sup>
12. Radical politics, usually New Left (cf. Baudrillard, Debord) Interdisciplinary scope and methods (though especially in linguistics, modern psychoanalysis, literary theory, and Strong sociology.) And, most pertinently: Writing style displaying high degree of abstraction, technicality, impersonality, and insularity. I should now try to account for what it is that "bad writing" is, but this is very outwith our scope. So, at the risk of making the endeavour circular, I will utilise the aesthetic-cum-moral criteria of pomophobes such as Denis Dutton<sup>12</sup>, in which "good" writing is identified with clear writing – the properties of being: quickly understandable even to nonspecialists; concrete (containing frequent reference to the 'ordinary' world); low on technical terms; lacking imagery; using short sentences which each represent one idea; and being unproblematic (perhaps I mean: "having one determinate sense").

The most common properties of "bad" academic writing are the negation of the above and the trio: difficult words (e.g. jargon, name-dropping, "isms", frequent

use of neologisms and foreign loanwords), difficult syntax (extended sentences), and difficult ideas (constant, nested abstractions).

It's spurious to examine works for "ease" outside of their proper context. However, a good example of what I think pomophobes have in mind is the 1998 winner in Denis Dutton's now-infamous Bad Writing contest<sup>13</sup>, Judith Butler's: The move from a structuralist account in which capital is understood to structure social relations in relatively homologous ways to a view of hegemony in which power relations are subject to repetition, convergence, and rearticulation brought the question of temporality into the thinking of structure, and marked a shift from a form of Althusserian theory that takes structural totalities as theoretical objects to one in which the insights into the contingent possibility of structure inaugurate a renewed conception of hegemony as bound up with the contingent sites and strategies of the rearticulation of power.<sup>14</sup>

Putting aside the question of the sentence's content: what reasons are there for this style?

#### Allegation #1: The French Deception

One suggestion is that massive, systematic obscurantism on the part of many thousands of people working in and around academia produces the style. That is, the claim that postmodernism en masse is intentionally and radically counterfeit. This is patently the lowest rank and grandest scale of conspiracy theory, and one propounded by people who characterize their enemies as the anti-rationalists. It chimes with the Classical smear-campaign that was run on the Sophists – another movement derided as rhetorical frauds.

#### Allegation #2: Keeping up with Profs Joneses

Alternatively, perhaps postmodernism is unintentionally elitist: perhaps a culture of bad rhetorical habits met with the ordinary pressures of self-interest, and led to confirmation bias and hot air. Practices might be reinforced when the peer community, worried about the need to impress, endorses each other even where discourse is turgid and/or hollow. This theory also seems quite silly, though is at least human-sounding. More sympathetically: academics are taught in jargon, and much of what we read is laden with it. We mimic this because distinctive jargon is customary to all fields. It is the convention that contains the other conventions.

Allegation #3: Argument from Sokal's hoax In 1994, a physicist, Alan Sokal, sent a deliberately absurd paper<sup>15</sup> to a cultural theory journal, Social Text; his parody passed for sincere deconstructive interdisciplinary theory and was published. Now, it is a good joke. But some use the affair as a reductio-ad-absurdum of all deconstruction, or all postmodernism, which it really is not. It is instead indication that one should not criticize from outside (that is, from a position of ignorance); that editors at this journal were lax; and that academic research is very probably vitiated by a process which does not require an expert in the discussed field to at least check for inaccuracy.

**Allegations #4 & 5:** The economics of thought A highly plausible, cynical pair of ideas: 1) That jargon proliferates because, where a coordinated group has an advantage, they will seek to cement that advantage. (i.e. it is in the interests of each brand of academics to have their own language, so that the long, expensive training they went through to become an insider is – is seen as – valuable).

- 2) That there is market pressure towards volume of published research (rather than “quality”, say), and that what wins journal space is what talks the orthodox contemporary talk.<sup>16</sup> Ease of comprehension is even interpreted as lack of quality in some contexts.<sup>17</sup> Jargon and the other hallmarks of research writing could thus be seen as a protective scholarly veneer of rigour and sophistication (or a “preference falsification”). DG Myers cites a philosopher feeling this moral hazard: “In the current crisis of hiring freezes and intense pressure for tenure, the need to publish is perhaps greater than any time before. Yet to publish in most journals means flinging the jargon, toeing the party line (which is somewhere to the left of gibberish), and quoting the usual suspects (Benjamin, Foucault, Derrida, Said, Jameson, Butler, etc.). I’m often appalled at my own writing, but since jargon, rather than substance, gains a publication, I succumb to verbiage.”<sup>18</sup>

**Justification #1:** The politics of clarity Some key linguistic assumptions of pomophobia are: that style and content are separable; that author-intention is the first and only really pertinent content of a text; and that “clear” writing is without ideological baggage. These claims have been dubious for some time, undermined in particular by the structuralist project.<sup>19</sup> From Roland Barthes’ defence of pluralism: “In truth, [clear] writing is clear only to the extent that it is generally accepted... For to write is already to organize the world, it is already to think.”<sup>20</sup> Texts are not to be thought of as divisible into form (language) and content (pre-language), because forms are themselves shot through with cultural assumptions. We should be wary of following “clear writing” on to “good writing”, since this is as much a socialized valuation as “good taste” or “propriety” have been.

Butler has defended her work from criticism by cruelly quoting Theodore Adorno, who “...surely had it right when he wrote about those who recirculate received opinion: ‘only what they do not need first to understand, they consider understandable; only the word coined by commerce (and really alienated) touches them as familiar.’”<sup>21</sup> His and her claim is that readable writing is so because it tends to be truism, the reuse of existing, politically corrupt ideas. The postmodern progressive sees a need to form “alternative procedures” of writing, and arguments often proceed from Marxist premises – for example, that what is “ordinary” or “common sense” is likely to be politically conservative (the theory of stable ideologies). From there they might conclude that the values and idioms of “clear writing” are basically bourgeois. “A piece is readable because familiar; familiar, because conventional; and conventional implies conservative.” There’s an idea, which even some philosophers (so-called Analytic philosophers like Stephen Stich and Brian Leiter) have taken up: “what is left for philosophy/the non-sciences

to do is to tidy up our thinking, no more”. Just resolve linguistic illusions; define; clarify; and maybe unify. Postmodernism is a countervailing tendency: the will to rip off the apron and throw out the dustpan.-From this perspective, a text’s being called problematic is not at all pejorative, but to be aimed for; jargon is taken to be a symptom of the struggle. Groundbreaking is messy, as it might be put clearly. Allegations #2, #4 and #5 all suggest processes that reinforce poor presentation habits – but these typically apply to all fields, and extend well beyond just academic discourses. None of the attributes of “bad writing” listed above are the sole preserve of postmodernists, nor are they notably characteristic. No one makes this criticism of Kant, and he more or less coined his own German to write in. Wittgenstein can often be whimsically unhelpful, as he concedes with stuff like: “I should not like my writing to spare other people the trouble of thinking.”

Justification #2: Insularity from specialization Jargon is also simply the shorthand of the professional, enabling concise, precise discussion amongst a pre-engaged group of peers. Catherine Belsey chides us: “[One reason we have difficulty reading Derrida] is that he is a Continental philosopher, with a range of reference that is not widely available outside that tradition. Many of his more impenetrable remarks turn out to be allusions to Plato, Hegel, or Heidegger, and not obscure at all to people who have those writers at their fingertips, in a way most of us don’t.”<sup>22</sup> I’m told that academic fields are fragmenting into subfields with their own niche journals; an academic today whose paper is read by a thousand people may be considered unusually successful. This addresses allegations #1 and #2: it is entirely more likely that there is no conspiracy at play, but simply that work is written to a committed, specialist audience. What this insularity says to the stated intent of many postmodernists (to effect change in the world) is another matter.

Justification #3: Intentional complexity or indeterminacy A friend of mine who tries to read Friedrich Nietzsche once described the ambiguous/symbolic style of Thus Spoke Zarathustra as “dickery” – but there is an underlying doctrine to it. Nietzsche wanted to emphasise the validity of perspectives and the fact that speech is never only one thing (the “multivocality” of language); this is writing made “difficult”, not so as to exclude, but to encourage multiple readings. This foreshadows the project which amounts to the positive element of postmodernism: it would not suit Derrida for us to ever fully understand his work, because this would assert our reading as The Reading, and this kills all the other ones. George Orwell notes that “Good prose is like a window pane”<sup>23</sup>, but this is precisely the opposite of the wall that postmodernists aim for. Some serendipity from Belsey: “... it is important from the point of view of the case against logocentrism to demonstrate in practice that language is not transparent, not a pane of glass through which ideas are perceptible in their pure intelligibility.”<sup>24</sup> Convolution makes the reader work. When reading Derrida, you are not allowed to forget that you are reading; the text’s opacity forces us to be reflective and reflexive, if not actually paranoid. It has been too easy to cry “Enlightenment!” in this discourse (for either side: as a rallying cry, or an insult). The debate cannot be rendered

simply as “linguistic transparency vs the ivory-tower” nor “radical intellects vs invested ideological absolutism.” These are biased strands of the whole braid of discourses, some of which are as old as rational inquiry: rhetoric vs reason; style vs/& content; action vs speech; the paternal elite vs the unsophisticated mass; truth vs/& imperialism; language vs/& thought; and the fundamental rift in political belief: tradition as font of wisdom vs tradition as the font of repression. Anyway, I refuse to take the criticisms above to be either the grand exposé or the pretension-puncturing service to humanity that pomophobes frame them to be. As an undergraduate (a hick) and a sympathetic sort (a mark), there’s a great deal to be done before I can write off Derrida et al. And if I eventually do, it will not be in the arena of style that the blade of my conclusion falls.

### <h3>Note on naming one's opponents</h3>

<div>

In the first version of this, written when I was a nipper, I used the word "pomophobia"

Some time later I read <a href="{{phob}}>this piece</a> by Nicholas Shackel, where he c

### <h3>Bibliography</h3>

<div>

<ul>

- <li>Anon (1997), Private communication to Denis Dutton, 15/6/1997</li>
- <li>Barthes, Roland (1966), <i>Criticism and Truth</i>, trans. Keuneman (London; Athlone, 1977)
- <li>Belsey, Catherine (2002), <i>Poststructuralism: A Very Short Introduction</i>; (Oxford, Oxford University Press, 2002)
- <li>Beardsworth, Richard (1996); <i>Derrida and the Political</i>; (Oxon, Routledge, 1996)
- <li>Benson, Ophelia & Stangroom, Jeremy (2006), <i>Why Truth Matters</i> (London, Cambridge University Press, 2006)
- <li>Bulhak, Andrew C (2000), <i>The Postmodernism Generator</i>, website</li>
- <li><http://www.elsewhere.org/pomo/></li>
- <li>Butler, Judith (1997), <i>"Further Reflections on the Conversations of Our Time" in Theory</i>, 1(1), 1-16
- <li>Butler, Judith (1999); <i>"A 'Bad' Writer Bites Back"</i>; op-ed in The Wall Street Journal, 12/12/99
- <li>Derrida, Jacques (1981), <i>Dissemination</i>, trans. Barbara Johnson (Chicago; London, University of Chicago Press, 1981)
- <li>Dutton, Denis (1998); <i>"Bad Writing Press Release, 1998"</i>, <http://www.denisdutton.com>
- <li>Fairclough, Norman (2001); <i>Language and Power</i> (Harlow; Pearson Education, 2001)
- <li>Foucault, Michel (1977); <i>"Nietzsche, Genealogy, History"</i>; in Language, Culture and Text, 1(1), 1-20
- <li>Galak, Jeff & Neilson, Leif (2010) <i>"The Virtues of Opaque Prose"</i>; Journal of Literary Theory, 14(1), 1-20
- <li>Lévinas, Emmanuel (1972); <i>Humanism of the Other</i>; (Chicago; University of Chicago Press, 1972)
- <li>Miller, James (2000); "<i>Is Bad Writing Necessary?</i>", Lingua Franca, vol.9 no. 1, 2000
- <li>Myers, D.G (2005), "Bad Writing", in <i>Theory's Empire: An Anthology of Dissent</i>, ed. S. M. Farber (London, Cambridge University Press, 2005)
- <li>Orwell, George (1953), "Why I Write", in <i>A Collection of Essays</i> (London, Penguin Books, 1953)
- <li>Sokal, Alan (1994), <i>"Transgressing the Boundaries: Towards a Transformative Political Poetics" in Cultural Studies</i>, 9(1), 1-20
- <li>Stolzenberg, Gabriel (2001), "Reading and Relativism: an introduction to the Scientific Study of Literature", in <i>Theory's Empire: An Anthology of Dissent</i>, ed. S. M. Farber (London, Cambridge University Press, 2005)

</ul>

</div>

# You do you

Gavin

2017-08-12

Out at London Pride, uncharacteristically. Around 1am, as you do, find myself in an intimate discussion with a stranger. Who leaps to what he imagines is my defence:

He: "What are you into?"

I: "Women, mostly."

"Oh right."

"It's a limited view, though. I wish I was more open."

"What? There's no such thing! You got to accept yourself!"

"Sure, but - I'm only attracted to like a tenth of the population; it would be objectively better to have a larger pool, and to see all of beauty."

"Hey, hey: stop it, you don't need to justify yourself. You are you, so you do you. Don't do anyone else."

"That's a beautiful thing to say, and there's a lot of people in the world who really need to hear it. Not me, though."

"It's true for everyone! You'll only hurt yourself by not being yourself."

"No, though. Consider: monkeys like bananas and sex; humans like bananas, sex, and philosophy and sports. So the value space visible to the latter is -"

"- Look, you're not going to change by overthinking things. So skip it! I'm not questioning you!"

"I am questioning myself. And so I often change. I don't want stock acceptance. Challenging me when I am not as good as I could be is better than acceptance: you take me seriously. You give me actual help."

"When I 'do' me, I justify myself to myself. In others this can be neurotic, self-sabotage. For me it is creation. If you wanted to google it, there's a word for it. Only at home when taking a pickaxe to the walls, to see if they're actually solid."

# Fair comparisons of cultural impact

Gavin

2017-02-10

Cultural footprints / Mindshare Hits / capita Hits / km<sup>2</sup> Hits / bn GDP

What year is your development environment  
from?

Gavin

2017-02-10

- Language (COBOL)
- Language version
- Version control
- Continuous integration
- Analytics

# The Ancient Greek industrial revolution

Gavin

2017-08-10

Greek “science” -> Stoics propositional logic -> computers The aeolipile -> steam power -> mining, naval domination Archimedes

Well behind in chemistry.

How much should we actually rate their theoretical? The Romans were better engineers, without much evidence of calculations

The economic collapse and the religious attack on intellect.

So what I’m saying is that if the two biggest events in European history - the fall of Rome and monopoly/totalitarian Christianity - hadn’t happened, THEN...

# Partisan verbs

Gavin

2017-10-07

We demand

We propose

# Qualitative features of programming languages

Gavin

2017-02-10

What I like in languages. - Low syntactic noise (Python and Scala over Javascript and Java. But then there's bash) - Hybridity (Scala over Haskell, PySpark over Pig. )

# Turing

Gavin

2017-09-01

```
{% assign cpi = "https://www.in2013dollars.com/uk/inflation/1940?amount=6000"
%} {% assign laiss = "https://www.panarchy.org/keynes/laissezfaire.1926.html"
%} {% assign ace = "https://en.wikipedia.org/wiki/Automatic_Computing_Engine"
%} {% assign cope = "https://sci-hub.se/10.1007/978-3-319-22156-4_3" %}
```

in the early days of computing, a number of terms for the practitioners of the field of computing were suggested in the Communications of the ACM — turingineer, turologist, flow-charts-man, applied meta-mathematician, and applied epistemologist.

- wiki

In a man of his type, one never knows what his mental processes are going to do next.

- JAK Ferns, Turing's coroner

*A review of Turing: The Enigma (1983) by Andrew Hodges.*

There have been two big films about Turing (three if you count the uselessly fictionalised Enigma (2001)). All are dishonestly melodramatic to some degree; for instance they depict Turing's relationship with his dead love Christopher as the driver of his work on machine intelligence. And more generally they depict him as tragic. But he wasn't tragic: we were. In the 1950s we attacked a superlative person, because we were certain it was the right thing to do.

Hodges, whose book began the great public rehabilitation of Turing and served as the source for the films, bears no blame for this: it's one of the best biographies I've ever read (better even than Kanigel on Ramanujan and Isaacson on Einstein). Hodges actually understands Turing's work, not just its consequences, and not just the drama around it. And what work!

```
{% include turing-bio/results.md %}
```

But even more than that: Copeland guesses that breaking U-boat Enigma saved 14 million lives, a large fraction of which we can lay at Turing's feet. If this is even roughly right this puts him in the top 50 life-savers ever.

But, outside of logic and engineering, where he was among the few most sophisticated people in the world, he was famously unsophisticated:

As at school, trivial examples of ‘eccentricity’ circulated in Bletchley circles. Near the beginning of June he would suffer from hay fever, which blinded him as he cycled to work, so he would use a gas mask to keep the pollen out, regardless of how he looked. The bicycle itself was unique, since it required the counting of revolutions until a certain bent spoke touched a certain link (rather like a cipher machine), when action would have to be taken to prevent the chain coming off. Alan had been delighted at having, as it were, deciphered the fault in the mechanism, which meant that he saved himself weeks of waiting for repairs, at a time when the bicycle had again become what it was when invented – the means of freedom. It also meant that no one else could ride it. He made a more explicit defence of his tea-mug (again irreplaceable, in wartime conditions) by attaching it with a combination lock to a Hut 8 radiator pipe. But it was picked, to tease him. Trousers held up by string, pyjama jacket under his sports coat – the stories, whether true or not, went the rounds. And now that he was in a position of authority, the nervousness of his manner was more open to comment. There was his voice, liable to stall in mid-sentence with a tense, high-pitched ‘Ah-ah-ah-ah-ah’ while he fished, his brain almost visibly labouring away, for the right expression, meanwhile preventing interruption. The word, when it came, might be an unexpected one, a homely analogy, slang expression, pun or wild scheme or rude suggestion accompanied with his machine-like laugh; bold but not with the coarseness of one who had seen it all and been disillusioned, but with the sharpness of one seeing it through strangely fresh eyes. ‘Schoolboyish’ was the only word they had for it. Once a personnel form came round the Huts, and some joker filled in for him, ‘Turing A.M. Age 21’, but others, including Joan, said it should be ‘Age 16’... It was demeaning, but the repetition of superficial anecdotes about his usually quite sensible solutions to life’s small challenges served the useful purpose of deflecting attention away from the more dangerous and difficult questions about what an Alan Turing might think about the world in which he lived. English ‘eccentricity’ served as a safety valve for those who doubted the general rules of society. More sensitive people at Bletchley were aware of layers of introspection and subtlety of manner that lay beneath the occasional funny stories. But perhaps he himself welcomed the chortling over his habits, which created a line of defence for himself, without a loss of integrity.

We have words for this now (“nerd”, “wonk”, “aspie”), and massive institutions, and even social movements, but at the time he had to make do with “don”, and hide inside academia. Again: the problem wasn’t him, it was us.

He gets called a mathematician most often, I suppose because people don’t want to be anachronistic. But scroll up: his most famous work is as a logician and a systems engineer, and the rest is statistics and algorithmics and cognitive science. He was falling between several chairs, until computer science caught up with him:

a pure mathematician worked in a symbolic world and not with things. The

machine seemed to be a contradiction... For Alan Turing personally, the machine was a symptom of something that could not be answered by mathematics alone. He was working within the central problems of classical number theory, and making a contribution to it, but this was not enough. The Turing machine, and the ordinal logics, formalising the workings of the mind; Wittgenstein's enquiries; the electric multiplier and now this concatenation of gear wheels – they all spoke of making some connection between the abstract and the physical. It was not science, not 'applied mathematics', but a sort of applied logic, something that had no name.

The philosopher-engineer. One of several moments in Hodge's book that left me dumbstruck is Turing arguing with Wittgenstein about the foundations of mathematics. (In the spring of 1939 they were both teaching courses at Cambridge called that!) Bit awkward, and in my view Alan goes easy on Ludwig. But you still couldn't make it up. The government employed Turing for 9 years, paying him about £6000 over the duration (£300k in today's money). In that time he produced 3 gigantically advanced systems (most of the Hut 8 system, the Delilah and the ACE design), about 10 or 20 years ahead of their time. Hodges sees this as a triumph of managerial socialism. Now, breaking naval enigma for £300k is an unbelievable deal (the savings from undestroyed shipping and cargo alone would be in the billions, let alone the loss of life, let alone the decisive tactical advantage). But the government suppressed Delilah and totally screwed up the ACE project. So I'm not sure if we can cheer too much. Keynes says somewhere that

The important thing for Government is not to do things which individuals are doing already, and to do them a little better or a little worse; but to do those things which at present are not done at all.

This is true of Bletchley. But instructive failures are only helpful if they occur in public. (As at least the ACE report was.)

The most annoying part of the films making up emotionally powerful unifying themes for Turing is that they are already there. But to grasp them, you'd have to actually display what was most wonderful and important about him, his technical work, and there goes the box office.

In an end-of-term sing-song [at Sherborne, when Turing was 12], the following couplet described him:

Turing's fond of the football field  
For geometric problems the touch-lines yield  
... another verse had him 'watching the daisies grow' during hockey... although intended as a joke against his dreamy passivity, there might have been a truth in the observation. [20 years later] ...One day he and Joan were lying on the Bletchley lawn looking at the daisies... Alan produced a fir cone from his pocket, on which the Fibonacci numbers could be traced rather clearly, but the same idea could also be taken to apply to the florets of the daisy flower. [30 years later] ...he was trying out on the computer the solution of the very

difficult differential equations that arose when [one] followed the chemical theory of [plant] morphogenesis beyond the moment of budding. . . . he also developed a purely descriptive theory of leaf-arrangement. . . using matrices to represent the winding of spirals of leaves or seeds round a stem or flower-head. . . The intention was that ultimately these two approaches would join up when he found a system of equations that would generate the Fibonacci patterns expressed by his matrices. . . Such observations reflected an insight gained from. . . [a program called] ‘Outline of Development of the Daisy’. He had quite literally been ‘watching the daisies grow’ . . . on his universal machine.

```
<h3>Highlights</h3>
<div>
    {%
        include turing-bio/quotes.html
    %}
</div>
<!-- -->
<h3>Why listen to me on this topic?</h3>
<div>
    <i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect
    <ol>
        <li>
            immersion in the field and/or good priors for what makes for an extraordinary claim in fiction
        </li><br>
        <li>
            incredible amounts of fact-checking gruntwork, at least 5x the time it takes to just read
        </li><br>
        <li>
            incredible amounts of argument-checking, which doesn't need domain knowledge.
        </li><br>
    </ol>
    I always try to do (3) but surely often fail.</i> <br><br><br>
```

```
    In this case: I am a computer scientist, and I've studied the early history of computing
```

*Cross-posted from Goodreads.*

# Did Turing commit suicide?

Gavin

2017-10-20

~ Alan Turing

The inquest into the death of Alan Turing was hasty and incompetent. For instance, the supposed suicide weapon, a poisoned apple, was never tested for poison. Turing's family ^

One great biographer, Andrew Hodges, considers it "obvious" that it was suicide.\* Another, Jack Copeland, argues that the evidence is more consistent with accidental poisoning.

Hodges is thorough: I got most of the counterevidence below from his book. But, among its other offences\*\* the film of his book invents wholesale a mental breakdown and teary confession of despair:

- "By the side of his bed was half an apple, out of which several bites had been taken. They did not analyse the apple, and so it was never properly established that, as seemed perfectly obvious, the apple had been dipped in the cyanide."

\*\* It invents a story about the police suspecting him of being a Soviet spy. It shows him covering up for a Soviet spy, John Cairncross; i.e. being a coward and traitor. It shows him building a computer with the same name as his first love, in an idiotic Frankenstein sense. It shows him having a breakdown after the hormone punishment. It hypes the relationship with Joan Clarke. It shows him being interrogated about his homosexuality, when actually he wrote a five-page confession outright, which is his largest contribution to gay rights: self-sacrifice as protest.

^ "John had already decided that it would be a mistake to contest a verdict of suicide, a policy from which the presence of a row of newspaper reporters did nothing to dissuade him."

This doesn't fit at all.

In the early 1950s, the suicide rate among British men in their early forties was 12 per 100,000. To my knowledge, there are no studies of suicide among gay men in the 50s, but in the 90s gay men in the US were roughly three times as likely to attempt suicide as straight men; this increase is a lower bound for the

1950s UK increase. The most relevant subpopulation, gay men convicted for having consensual sex and put on diethylstilbestrol, is a blank page.

The rate of death by accidental poisoning in this period is about 90 per 100,000. Having been found dead, the probability of the cause being suicide was about 1% for a man like Turing (41 in 1954).

The suicide rate would be higher for convicted people who have been castrated. But so would it be for “slovenly people who have cyanide experiments evaporating in the next room”. This might even out, I don’t know.

`suicide given death: 1%`

`suicide given ...`

`The way to get a probability for an event which only happens once is to use Bayesian updatin`

Evidence for suicide

2 years earlier, unjust conviction and medical torture.

20 years earlier, he mentioned a design for a suicide to a friend, involving an apple and electrified wire.<sup>^^</sup>

3 months earlier, made a new will. (“There was one piece of evidence that he had prepared for death: he had made a new will on 11 February 1954. This in itself was a kind of statement as to where he stood.”)

Testimony of Allan Pacey, an expert in male fertility that the effects of stilb are permanent.

letters to Robin Gandy talks about his “shocking tendency at present to fritter my time away in anything but what I ought to be doing.”

Fond of Snow White

- in mid-May 1954, when Alan went with the Greenbaum family for a Sunday visit to Blackpool. It was a very fine day, and they walked cheerfully along the Golden Mile of seaside amusements, until they came across the Gypsy Queen, the fortune-teller. Alan went in, to consult her. Had not a gypsy foretold his genius, in 1922? The Greenbaums waited outside, and found themselves waiting for half an hour. When he came out, he was as white as a sheet, and would not speak another word as they went back to Manchester on the bus. They did not hear from him again, until he called on the Saturday two days before his death when, as it happened, they were out. They heard of his death before returning the call.

<sup>^^</sup> “James [Atkins] also had a letter from Alan which described, in a rather detached way, that he had been feeling depressed and mentioned that he had even thought of a scheme for ending his life. It involved an apple and electrical wiring.”

<sup>^</sup> Typical error: ‘However the coroner’s report, also on display, is unequivocal: Turing had consumed the equivalent of a wine glass of poison.’ It actually says that

his stomach contained four ounces of bitter-smelling fluid, which inhalation could also easily cause. <https://www.theguardian.com/science/2012/jun/20/alan-turing-science-museum-exhibition>

#### Evidence against suicide

- Throughout his life, he was sloppy with hygiene and tidiness. (“The electrolysis experiment was wired into the ceiling light socket. On another occasion, an experiment had resulted in severe electric shocks. And he was known for tasting chemicals to identify them.”)
- Autopsy shows slow. No contortion. Ingestion is rapidly and violently lethal, seconds.
- No mention to psychiatrist or closest confidants, both of whom he told the deepest personal problems to.
- “The day of the trial was by no means disagreeable. Whilst in custody with the other criminals, I had a very agreeable sense of irresponsibility, rather like being back at school.”
- Turing had cyanide in his house for chemical experiments he conducted in his tiny spare room - “the nightmare room”.
- No evidence of pre-meditation. This leaves 1) intentional hiding or 2) impulsivity. What % of suicides are apparently impulsive?

Or perhaps, more likely, he had accidentally inhaled cyanide vapours from the bubbling liquid. The nightmare room had a “strong smell” of cyanide after Turing’s death. (cyanide inhalation leads to a slower death than ingestion and wouldn’t necessarily induce contorted death throes)

- there was no clear pattern of decline or failure in his intellectual life that might in itself explain its abrupt end. It was rather a fluid, transitional period such as had occurred before in his development, and this time accompanied by a wider range of interests, and a more open attitude to intellectual and emotional life.
- It fell into no clear sequence of events. Nothing was explicit – there was no warning, no note of explanation. It seemed an isolated act of self-annihilation.

#### good health freedom from financial trouble

- while Alan lay in his small front bedroom, an electrolytic experiment was bubbling away at the back... He did sometimes use [potassium] cyanide for electrolysis, it being necessary for gold-plating... She argued that he had got cyanide on to his hands by accident, and thence into his mouth. This was, of course, what she had always said might happen. At Christmas 1953, when he made his last Guildford visit, she had repeated her warning ('Wash your hands, Alan, and get your nails clean. And don't put your fingers in your mouth!').

- His working papers were left in an untidy mess in his room at the university... there was no wholesale clearing up, neither of personal papers nor of his research. It was as though he had planned for the possibility, but in the event acted impulsively.
- He had also booked as usual to use the computer on the Tuesday evening, and the engineers waited up for him, only hearing next day that he was dead.
- His friendly next door neighbours, the Webbs, had moved to Styal on the Thursday, and he had had them to dinner on the previous Tuesday, merry and chatty. He had been much regretting their move, spoke of visiting them, and said he was glad that the new occupants would be young and with young children.
- There were purchases, including theatre tickets, in his house when he died; he had written, though not posted, an acceptance of an invitation to a Royal Society function on 24 June.
- “Neither, indeed, did his Guildford visit [to his mother] take on a leavetaking character”
- Robin Gandy: Turing “seemed, if anything, happier than usual”.

Evidence neither way

o There was a half-eaten apple beside his bed. However, he usually ate an apple before bed. (At wartime Hanslope Park, “Alan was also able to have the apple that as a rule he would always eat before going to bed.”)

Given the weakness of the evidence, why is the suicide hypothesis the received wisdom? Well, without over-egging it: it makes a better story. And it gives the event meaning - terrible, but morally straightforward meaning. It makes Turing’s death about homophobia, rather than an absurd empty accident.

# Modelling linguistic accommodation

Gavin

2015-07-01

{% include accommodation/links.html %}

0

For example, you may come across definitions like this: “A finite state automaton is a quintuple (

$Q$

,

$\Sigma$

,

$q_0$

,

$F$

,

$\delta$

) where  $Q$  is a finite set of states (

$q_0$

,

$q_1$

, . . . ,

$q_n$

),

$\Sigma$

is a finite alphabet of input symbols,

$q_0$

is the start state,

$F$

is the set of final states

$$F \in Q$$

, and

$$\delta \in Q \times \Sigma \times Q$$

, the transition function.”

That definition should be taken outside and shot. 3

~ John Coleman

‘Accommodation’ is that thing where you automatically mimic the person you’re talking to. You might immediately think of baby talk and speaking loudly to old people, but these conscious games are not what I’m talking about. For humans also unconsciously shift speech, depending on the gender, status, and likeability of their interlocutor. Accommodation is pervasive, correlated with key bits of human interaction: empathy, status, and teamwork. Studying accommodation puts you at the intersection of statistical modelling, linguistics and social signal processing.

The trick is to detect it using Hidden Markov models (HMMs). A technique from speaker verification is adapted: model-conditional probabilities estimate the ‘distance’ of each speaker, from their interlocutor, for each word. This likelihood ratio is taken for each word uttered by a speaker, relative to their interlocutor uttering it. The correlation of these ratios over time is used to infer the presence of accommodation and estimate effect sizes.

I used the dataset from Stuart-Smith et al (2015) (henceforth “SSSV” after the surnames of the authors): n=120,000 words, from 6 pairs of speakers. The modelling and data analysis was implemented in Python, with modelling tools from the ‘Hidden Markov Toolkit’ (HTK). 2

## 1. Introduction

{% include accommodation/intro.md %}

## 2. Glossaries

{% include accommodation/gloss.md %} {% include accommodation/maths.md %}

## 3. Sequence modelling with state machines.

{% include accommodation/fsm.md %}

## 4. Hidden Markov modelling for linguistics

{% include accommodation/comp-ling.md %}

## 5. Present methodology

{% include accommodation/method.md %}

## 6. Results

```
{% include accommodation/results.md %}
```

## 7. Conclusion

```
{% include accommodation/conc.md %}
```

```
{% include accommodation/biblio.html %}
```

```
{% include comments.html %}
```

```
{% include accommodation/foots.html %}
```

# Is modelling individuals good?

Gavin

2017-10-20

Cathy O'Neil does a great service in rendering clear the doubly obscure world of big-data modelling. (Doubly: both highly technical and private to corporation's internal code.)

But she doesn't do anything like a cost-benefit account of the practice. She likes FICO, because it had good effects on loans for underrepresented people. It did this by allowing better modelling.

Is Risk Classification Good for Society?

From a societal perspective, is risk classification desirable?

Some argue that risk classification should be restricted when

insurance is mandatory (e.g., auto liability)

classification is based on inherited traits (e.g., gender, genes)

classification is based on location of residence (e.g., auto, property)

classification is based on subjective criteria (e.g., "poor moral risks")

Effects of restricting classification

1. Redistributions income

From low risk to high risk

Is this fair?

2. Changes behaviour

Classification will alter insurance prices to certain groups and therefore change behavior

Types of behavior:

amount of insurance purchased

loss control activities

Some changes in behavior may be desirable and some undesirable

amount of liability insurance purchased by poor people

smoking

amount of life insurance purchased by people with HIV

3. Decreases classification social cost

Ignoring fairness issues (point #1), if there are no behavioral effects of classification  
i.e. classification simply redistributes income

4. Limiting classification may increase regulatory social cost

Monitoring to enforce restrictions

Need to impose other costly restrictions on insurers

marketing activities

underwriting activities

Restrictions lead insurers to not offer coverage

Leads to residual market (involuntary market) mechanisms

Leads to additional costs

# Conceptual conversions

Gavin

2015-02-06

```
{% include conversion/style.html %}  
{% include conversion/main_table.html %}  
Tenuous  
{%   include conversion/tenuously.html      %}  
{% include conversion/foots.html %}
```

# How lethal are the Tories? Part 2

Gavin

2017-09-20

## **Part 2:**

<https://www.ncbi.nlm.nih.gov/pubmed/28208027> [http://www.dannydorling.org/?page\\_id=5942](http://www.dannydorling.org/?page_id=5942)  
<https://www.rsm.ac.uk/about-us/media-information/2017-media-releases/new-analysis-links-30000-excess-deaths-in-2015-to-cuts-in-health-and-social-care.aspx> <http://www.geog.ox.ac.uk/news/articles/170217-ddorling-cuts-excess-deaths.html>

# How false is TED?

Gavin

2017-08-15

Meta-meta-analysis.

Compared to social science's average replication rate (50%)

# Why I'm not a philosopher

Gavin

2017-08-20

```
{% include nophil/links.md %}

<h3>Resolution (2021)</h3>
<div>
    I have been flip-flopping on this post every few months for 4 years. I've cracked it at
    Is the <a href="{{koko}}>left tail</a> (Herder, Rousseau, Marx, Freud, Zhu Hongdeng, Inazō
        <center><a href="#fn:1" id="fnref:1">1</a></center>
        <br><br>

    I think it's extremely difficult to know your own potential, and also surprisingly diffi
        So my actual answer to the statement in the title is: because I am not confident I'm ex
</div>
```

Can you tell them, with a straight face, to follow philosophical argument wherever it may lead? If they challenge your credentials, will you boast of philosophy's other great discoveries: that motion is impossible, that a Being than which no greater can be conceived cannot be conceived not to exist, that it is unthinkable that anything exists outside the mind, that time is unreal, that no theory has ever been made at all probable by evidence (but on the other hand that an empirically adequate ideal theory cannot possibly be false), that it is a wide-open scientific question whether anyone has ever believed anything, and so on, and on, ad nauseum? Not me!

– David Lewis

People are not confident [analytic philosophy] can solve its own problems, not confident that it can be modified so as to do better on that first score, and not confident its problems are worth solving in the first place... what we see is a desperate scramble to show that the skills or tools we have might find some problem space wherein their, our, worth can be made manifest... I do not think such a problem space has been forthcoming.

– Liam Bright

It's simple: The greatest nontechnical minds in history have all failed to work out the nature of the world just by thinking about it, and so would I.

(Technical minds sometimes manage it, but only with a lot of help from data, plus maths, plus just thinking about it. But that isn't philosophy, anymore.)

...

Alright alright it's *not* simple. Aside from the pursuit of truth, which it is manifestly bad at: why do philosophy?

```
<li>
    <h3>the philosopher as intellectual janitor</h3>
The standard rejoinder to the account of philosophy implied above is that philosophy is not
<br><br>

    <blockquote>
        ...it is scrutiny of [the] uncritical acceptance of the realm of physical objects itself
    </blockquote>
<br>

Or Wittgenstein, the radical janitor:
    "<i>In philosophy we are not laying foundations but tidying a room, in the process of which
<br><br>

The standard rejoinder to this rejoinder is to ask for <a href="{{hue}}>a single natural law</a>
</li>
<br><br>

<li>
    <h3>philosophy as justification of belief and action</h3>
Maybe philosophy's job is giving a general "foundation" to what we do. That is, it doesn't do
    <blockquote>
        Throughout my writings I have made it clear that my method imitates that of the architect
    </blockquote>

I think it's fair to regard this as a dead-end: after thousands of years of trying, we appear
More controversially, I'm no longer sure why we need it. Many things don't seem to need (phi
<br><br>

If you pointed a gun at me, I'd answer with some blend of pragmatism ("whatever works is just
</li>
<br><br>

<li>
    <h3>philosophy as activity</h3>
Another common one is that philosophy isn't a thing (e.g. a body of claims), but a process,
    <blockquote>The crux is what <span style="font-weight:bold">happens</span> in [philosophy],
    </blockquote>
Or Fichte:
    <blockquote>
        Make no mistake about this: nothing that I or any other teacher can lecture to you about
    </blockquote>
```

(This <a href="{{grace}}>explains why we read so much old/obsolete work</a>: we're learning  
</li>  
<br><br>

<ul>  
<li>  
    <h4>philosophy as virtuous self-examination</h4>  
    A <a href="{{delp}}>literally</a> <a href="{{gnothi}}>classic</a> view is that philosophy  
    <!-- -->  
    Or, more recently, Alain de Botton: <br><br>  
    <blockquote>  
        Socrates compared living without thinking systematically to practicing... [e.g. pottery]</blockquote>

But you can't understand yourself if you're not right about yourself. Nor can you be 'authentic'.  
<a href="{{stove}}>strange false inferences</a> great philosophers have made about themselves.  
<!-- In general, introspection is overvalued as a source of deep psychological truths (related to</li>  
<br><br>

<li>  
    <h4>philosophy as therapy for the human condition</h4>  
    Another ancient claim: philosophy is good for your mental or spiritual health - for instance,  
<br><br>

Epictetus: <br><br>  
<blockquote>A philosopher's school is a surgery: pain, not pleasure, you should have felt there</blockquote>  
Or again Alain de Botton:<br><br>  
<blockquote>art and philosophy help us... to turn pain into knowledge.</blockquote>

However, despite this long tradition, whether philosophising leads to peace of mind is an empirical question.  
<br><br>

Anecdotes abound. Plenty of people say that Stoic philosophy made their life better. But the evidence is mixed.  
<br><br>

Real Buddhist practice <a href="{{budd}}>seems to run similar risks</a> of permanent dissatisfaction.  
</li>  
<br><br>

<ul>

<li><h3>philosophy as <a href="{{sss}}>state space search</a> over coherent worldviews</h3>  
Maybe philosophy doesn't have to answer questions to be useful. We can read Cicero's ancient  
<br><br><blockquote>

    There is nothing so absurd that it has not been said by some philosopher.  
</blockquote>

as a compliment: we consider everything. Philosophers are then in the business of conditionals.  
<br><br><blockquote>

Is scientific progress useful for philosophy? Certainly. The realities that are discovered

</blockquote>

and <a href="#">[Massimo Pigliucci](#) :

<br><br><blockquote>

Unlike science, where we do seek answers to questions determined by empirical evidence,

</blockquote>

A priori reasoning can't tell you how the world is, but it can tell you what the world [is](#).

<br><br>

We need more truths first, to help control the combinatorial explosion of possible philosophies.

decades of fruitless Unified Field work</a>, or with hundreds of <a href="#">[weak models](#)

</li>

<br><br>

<li>

<h3>what about <a href="#">[experimental philosophy](#) ?</h3>

They've got the right idea: they don't rely solely on intuition and deduction. Millions of

<br><br>

But the x-phi people aren't doing philosophy in the bit of their work that is distinctive. They

<br><br>

Related: there are of course hybrid scientist/philosophers, with more hope. The most important

<br><br>

This gives the game away again: it is really only apriorism I'm disparaging, the idea that

</li>

<br><br>

<li>

<h3>what about logic?</h3>

The logic department get a lot of objective, objectively important stuff done. And there are

<br><br>

But their methods are quite far from the core of the field; they are castaways of a <a href="#">historical accident; their closest kin are in computer science or maths departments. (All

<a href="#">[programming](#), for instance.) Logic is a member of <a href="#">[that](#)

</li>

<br><br>

<li><h3>philosophy as improving us for other enquiry</h3>

Bertrand Russell: <br><br>

<blockquote>

"Philosophy is to be studied, not for the sake of any definite answers to its questions

</blockquote>

Does the study of philosophical questions actually make us better scientists or citizens?

<br><br>

Even if it does, this still implies that philosophy is secondary: that one should use philosophy

</li>  
<br><br>

<li>  
    <h3>philosophy as improving our view of what the world <i>should</i> be</h3>  
Some ethical philosophies don't aim at discovering truths, and yet (a handful of) ethicists  
Yes: there has been moral progress, and some of this is due to philosophy. (I know this, because  
<a href="{{dave2}}">David Pearce</a>'s essays changed my life, and they are half conventional  
<br><br>

Even then, I think the expected value of being an average ethicist is probably less than that.  
</li>  
<br><br>

<li>  
    <h3>philosophy as about truth - we just haven't had enough time</h3>  
There are <a href="{{greg}}">hundreds of times more</a> philosophers working now than in past  
<br><br>

The clock time spent on philosophy is impressive: 3000 years. But the above implies that the  
Sure, the distribution of philosophical workers is skewed towards the present and future. But  
<br><br>  
<!-- -->  
Our sample size isn't very large for some subfields. My favourite research programmes are <a href="#">  
<br><br>

Other reasons philosophers today should be the best:

<ul><li> Actual constraints on reality from fundamental physics.</li>  
<li> Powerful logics (FOL, HOLs, modal, utility theory)</li>  
<li> Free library of almost every other philosopher ever, most of whom speak the same language  
<li> Computers for <a href="{{quant}}">simulation</a> & note-taking & word processing even</li>  
</ul><br>  
(Constraints make it easier to find the truth, but harder to publish arbitrary things.)<br></li>  
<br>

<li>  
    <h3>philosophy as defence against unavoidable philosophy</h3>  
    Maybe you either do philosophy explicitly, or get pwned by a bad (or anyway unvetted) philosopher  
<blockquote>  
        What we are destroying is nothing but houses of cards and we are clearing up the ground of  
</blockquote>  
I reject this view because I don't think philosophical problems generally are just linguistic  
<a href="{{dave}}">David Pearce</a>, a scientifically literate philosopher, believes this:<br>  
<blockquote>  
    The penalty for not doing philosophy isn't to transcend it, but simply to give bad philosophy  
</blockquote>

```

True, but you don't need to be a philosopher to watch out for sneaky philosophers.
</li>
<br><br>

<li>
    <h3>philosophy as fun</h3>
    What if I just really like it? Like Hume: <br><br>
    <blockquote>
        I cannot forbear having a curiosity to be acquainted with the principles of moral good and evil.
    </blockquote>
    This is also fair enough, except that I think we have a duty to do more than please ourselves.
    <br><br>

    This kind of philosophy is a game - the hardest game, yes, since <a href="{{nom}}">
    the rules</a> <a href="{{dial}}">are themselves</a> at issue. I love it, but that is not enough.
    </li>
    <br><br>

    <li>
        <h3>the present work's sceptical empiricism as philosophy</h3>
        <center>
            <blockquote>
                To mock philosophy is to be a true philosopher.
            </blockquote>
            - Pascal
        </center>
        <br>
        Isn't this essay a work of (meta)philosophy, and am I not drawing serious, useful inference?
        <!--  -->
        Well, <a href="#point">my original point</a> was an induction from past philosophy to my
    </li>

```

Where does this leave us?

I find myself piling up many kinds of philosophy one should be doing - e.g. negative philosophy against bad philosophy, practical ethics, schemes for handling moral uncertainty, logic, population ethics, existential risk. But then I remember the left tail, of very harmful philosophers.

The relatively small active effort on many questions (at most a few hundred careers, and more often much less than one) is a good argument for it not being impossible to solve philosophical questions. (Less-likely impossible in proportion to neglect.) Also there's the importance of non-perverse philosophy for making a future artificial intelligence; it doesn't need to be right or definite, but it needs to land in a non-insane part of philosophical space.

A real nonphilosopher would not feel the need to write something like this.

*Dedicated to the University of Aberdeen, who in a 6 year period either fired, lost,*

*or pushed out of teaching all but one of the excellent philosophers who taught me:* Gerald, Joe, Bob, Gerry, Guido, Nate, Tony, Catherine, Crispin, Grant, Russell, Aidan, Dylan, Aaron, Filippo, Francesco, Luca.

## See also

```
{% assign schw = "https://schwitzsplinters.blogspot.com/2020/08/philosophy-that-closes-vs-philosophy.html" %} {% assign stove = "https://web.maths.unsw.edu.au/~jim/wrongthoughts.html" %} {% assign spol = "https://www.jelonsoftware.com/2001/04/21/dont-let-architecture-astronauts-scare-you/" %}
```

- Broadness as trivial predictor of philosophical status
- Tom Adamczewski, Philosophical Success Stories
- Massimo Pigliucci, Progress in Philosophy
- Graham Johnson, Conceptual engineering: the revolution in philosophy you've never heard of
- Schwitzgebel on philosophy that opens
- The dangerous ravings of David Stove
- Luke Muehlhauser, Philosophy: A Diseased Discipline
- Spolsky on the risks of abstraction even in practical matters
- David Pearce, Long diary entry containing a metaphilosophy & applied physicalism.

```
{% include nophil/tangents.html %}
```

Of course, the tails of science are also heavy. Heavier. But it's easier to tell which you're in.

# Every Mountain Goats Song that has Been Recorded and Does Not Have a Death Geas Placed Upon It By Its Author

Gavin

2017-12-31

“Tracking live is like that. You’re actually hearing something which actually happened.”

Many of my tasteful friends disdain even his soft hi-fi work, and some of my tasteless friends are much enamoured of the lot. Just like anyone, I cannot rule out ‘adaptive preferences’, that I like it because I’ve put a lot into it.

There’s just a lot to it. He has a lot to say, a lot to remember. He does a dozen interviews a year, and I watch em all, and he remains full of new stories, recommendations, enthusiasms and disgusts.

B sides, deep cuts, no cuts.

Major-key desolation

Completeness like this post is out of keeping with the man himself. Darnielle likes lost arcana. (This was a joke: “Plenty of people take up painting; me, I like to hand-make little booklets and tuck them away in a shoebox where they’ll never be seen again. There are twenty-two such booklets so far, and they’re some of the best work I’ve ever done, and my plan is to bury them in the backyard when I’ve finished with them. In partial shade. Near the raspberry bushes.” But it is not outside the realm of possibility.)

And his live shows are exciting because there is much old treasure, quite beside the new, quite besides the stories, and quite besides his joy.

The autobiographical ones can be split into childhood, vampires (drug squatting), and .

The hoarse yell of deviance

His latter albums, what with less screaming, are only boring if you don’t know the pre-2002 holler; the careful mutter of Life of the World is absolutely an art choice.

My album ranking is mostly a function of the constituent songs' ranking. Though some work together to greater effect, like Tallahassee or MCB.

All of these songs can receive a +3 at a given live performance. The trouble is that between AHWT and TLotWTC he became a musician. I love the demos of the recent smooth albums. There only used to be demos.

- He has destroyed about 50
- And about 10 have only been played live
- And several exist ambiguously In title only Perhaps
- A new track was found last year on one copy of an early cassette
- And no other copy of that cassette
- Sometimes We Mosh - live only, lyrics here.

# A ceiling for human expertise

Gavin

2018-01-31

{% include ceiling/links.html %}

This figure appears in DeepMind's instant-classic paper 'Mastering the Game of Go without Human Knowledge' (2017):

Figure 3b: 'Prediction accuracy on human professional moves. The plot shows the accuracy of the neural network at each iteration of self-play, in predicting human professional moves... The accuracy measures the percentage of positions in which the neural network assigns the highest probability to the human move.'

It shows that AlphaGo Zero (AGZ) only predicts human pro moves with 50% accuracy, at best. That is, AGZ disagrees with human professionals on 50% of moves.

This perhaps has implications for human expertise in general, by the following argument:

1. AGZ plays far beyond peak human ability.
2. AGZ would play differently from a peak human in 50% of moves.
3. So a peak human makes suboptimal moves at least 50% of the time.
4. Go is an excellent environment for human learning  
(small ruleset, rapid objective feedback, amenable to intuition).
5. So, relative to more complex domains, human mastery of Go should be relatively complete.
6. So we can expect human experts in other, more complex domains to make suboptimal decisions at least 50% of the time.

Regarding premise 4, Ericsson says learning occurs if people are "1) given a task with a well-defined goal, 2) motivated to improve, 3) provided with feedback, 4) provided with ample opportunities for repetition and gradual refinements of their performance"

# ‘The Great Influenza’ (2004) by Barry

Gavin

2018-06-26

{% include flu/links.md %}

“It seems to be a plague, something out of the middle ages. Did you ever see so many funerals, ever?”

– Catherine Ann Porter

A rousing history of one of the worst things to ever happen: the 1918 outbreak of H1N1 flu. Most of it focusses on the frantic research against it; I’d never heard of any of the scientists. They didn’t win, but they got us ready for next time.

Barry senses that the headline result - one-third of the entire world infected, with 25-100 million dead - is a numbing number. So, in modern terms:

It killed more people in twenty-four weeks than AIDS has killed in twenty-four years, more in a year than the Black Death killed in a century.

Or ten thousand 9/11s. It’s worth belabouring this, because we have a terrible habit of paying far more attention to human threats than natural ones, even when natural ones are far worse. (Witness our terrorism prevention budgets compared to our infectious disease control budgets, when the latter is a thousand times more lethal.)

So: The 1918 flu was worse than the entire First World War: 40+ million died of flu, compared with 17 million dead from war. 500 million people were permanently damaged by flu, vs 41 million by the war. 3% of all humans died of flu, including about 8% of young adults!.

But it’s hard to separate the War and the pandemic. The virus was spread everywhere by unprecedented numbers of troops, and by the massive supply convoys it induced, and by the War’s other human displacements. We don’t know how many of the pneumonia deaths only occurred because of the logistical degradation, poverty and pestilence of wartime. There are terrible nonlinearities involved in overcrowding and global movement of troops. But add millions at least to the overall death toll caused by WWI.

## **Therapeutic nihilism**

The first third of the book is a prelude, describing how terrible medicine was up to the 20th Century. Medicine was “the withered arm of science”. Therapeutic nihilism (that is, “we can’t really do anything”) was the rational view, replacing millenia of Galenic woo.

Stengel reviewed dozens of ideas [for H1N1 treatments] advanced in medical journals. Gargles of various disinfectants. Drugs. Immune sera. Typhoid vaccine. Diphtheria antitoxin. But Stengel’s message was simple: This doesn’t work. That doesn’t work. Nothing worked... Nothing they were doing worked.

But this created a powerful vacuum: humans want to believe something can heal. The gap was filled with worse. Confabulations from this time still haunt us: homeopathy, chiropractic, naturopathy, Christian Science, and (though Barry doesn’t include them) the organic farming movement and psychoanalysis.

Few people come off well. Even among the scientists, we get a horrible example of perverse priors and premature updating: most scientific resources were devoted to fighting the wrong pathogen, due to a stubborn bad guess by an extremely eminent researcher.

<h3>Rockefeller Institute</h3>

<div>

Quite a lot of the entire world's research funding for H1N1 was concentrated in the <a href="#">Rockefeller Institute.

They'd make for a good case study in ultra-effective philanthropy, though of course in this case it was probably not intentional.

</div>

## **War: reportedly hazardous to public health**

There is, therefore, but one response possible from us: Force! force to the utmost, force without stint or limit, the righteous and triumphant force which shall make right the law of the world and cast every selfish dominion down in the dust.

– Woodrow Wilson addressing one of his infective money-lending mobs.

Wilson tends to be viewed pretty positively, because he won. (“at last the world knows America as the savior of the world!”) But in the process he perverted an entire state and nation; ignored the terrible suffering of his own population for years; and refused a conditional peace with Austria in August, and again with the Kaiser’s new parliament in September. (This meant 70 extra days of war, which, if this period was as lethal as the rest of the war, means up to 800,000 completely unnecessary deaths, not counting the collateral damage from wasting even more medical resources, mixing the population even more, during the worst epidemic ever).

the military suetioned more and more nurses and physicians into cantonments, aboard ships, into France, until it had extracted nearly all the best young physicians. Medical care for civilians deteriorated rapidly. The doctors who

remained in civilian life were largely either incompetent young ones or those over forty-five years of age, the vast majority of whom had been trained in the old ways of medicine.

He did great harm and should be viewed as we view Wilhelm II, whatever his unconsummated ideals. And this is before we consider blaming him, or the bloody virus, for the Treaty of Versailles, and so the rise of the Nazis.

But on April 3, 1919, Wilson fell ill with flu-like symptoms. . . Ever since, historians have wondered about this episode, both concerning Wilson's prior health problems and his performance when he returned to the negotiating table a week later. Wilson wasn't the same man. He tired easily and quickly lost focus and patience. He seemed paranoid, worried about being spied upon by housemaids. He achieved some of his specific goals but was unable or unwilling to articulate a broader vision for a better world. In other words, he acted like a man with residual neurological problems stemming from a recent bout of Spanish flu. Over the next crucial weeks, Wilson lost his best chance to win the peace by agreeing in principle to draconian terms favoured by France. The final settlement punished Germany with a formal admission of guilt, enormous reparations and the loss of about 10 per cent of its territory.

This is too neat, too terrible. It reads like greentext, though all of the steps make sense (H1N1 cases in his entourage; severe cognitive deficits from recovered patients). Wikipedia doesn't even mention it, so I suppose it's fringe. Barry is aware of the temptation to tie everything into one knot, and hedges.

You already believe, probably, that World War I was a terrible senseless waste of life. Well, now magnify that belief by a factor of 5 or 6.

```
{% include flu/crimes.md      %}  
{% include flu/unamerican.md %}
```

### The undocumented apocalypse

Because the disease was everywhere, ravaging the species (and beyond), the book can't cover everything. Very little is said about non-Americans, i.e about 98% of the death and chaos. This is partly because there just isn't a lot of evidence about them, despite their influenza immunity and medical care being even worse. (This is why the top estimates reach 100m deaths, three times the median estimate.)

Here is a passage about just a tiny number of them, in the north:

In Alaska, whites protected themselves. Sentries guarded all trails, and every person entering the city was quarantined for five days. Eskimos had no such luck. A senior Red Cross official warned that without "immediate medical assistance the race" could become "extinct." . . . The navy provided the collier USS Brutus

to carry a relief expedition. . . They found terrible things. One doctor visited ten tiny villages and found “three wiped out entirely; others average 85% deaths. . . Survivors generally children. . . probably 25% frozen to death before help arrived.” The virus probably did not kill all of them directly. But it struck so suddenly it left no one well enough to care for any others, no one to get food, no one to get water. And those who could have survived, surrounded by bodies, bodies of people they loved, might well have preferred to go where their family had gone, might well have wanted to no longer be alone. . . Two hundred sixty-six people had lived in Okak, and many dogs, dogs nearly wild. When the virus came, it struck so hard so fast people could not care for themselves or feed the dogs. The dogs grew hungry, crazed with hunger, devoured each other, then wildly smashed through windows and doors, and fed. . . In all of Labrador, at least one-third the total population died.

{% include flu/phill.md %} {% include flu/err.md %}

*Cross-posted from Goodreads.*

# Intro to virtue epistemology

Gavin

2012-02-14

```
include virtueep/links.html
{%
  include virtueep/glossary.html
}%}
```

# Notes on AI Safety

Gavin

2018-03-28

```
{% include safety/links.html %}  
{% include safety/motivation.html %}  
Technical  
{%   include safety/armstrong-IRL.html      %}  
{%   include safety/worley.html      %}  
Policy  
{%   include safety/brundage-avin.html      %}  
Dissent  
<ul>  
  <li>Kaufman</li>  
  <li>Hanson</li>  
  <li>Reese</li>  
  <li>Chapman</li>  
  <li>Pinker</li>
```

# Neither Turing, neither Searle

Gavin

2009-03-01

{% assign crux = "https://en.wikipedia.org/wiki/Experimentum\_crucis" %} {%  
assign discourse = "https://www.gutenberg.org/files/59/59-h/59-h.htm" %}

SIMPLICIO: ‘Some computer programs might be able to pass a Turing test, but that doesn’t provide any evidence that they can think. They might use all the right words, but that doesn’t mean they understand what the words mean.’

The Turing test is sometimes portrayed as a proper crucial experiment verifying the presence of intelligence - i.e. a sufficient condition for thought - and sometimes just as evidence for thought. But it was actually originally intended to *sidestep* the question of whether machines can think: Turing deemed that “too meaningless for discussion.”<sup>1</sup> His replacement question is:

Is it possible for a finite-state digital computer, provided with a large... program, to provide responses to questions that would fool an unknowing interrogator into thinking it is a human being?

(In fact Turing made a precise forecast, specifying the memory bounds, and a point estimate of when it would be passed with specific accuracy:

I believe that in about fifty years’ time it will be possible to programme computers, with a storage capacity of about 10<sup>9</sup> [bits], to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning.

This forecast did not come to pass (and still hasn’t after 73 years), despite ordinary computers now having more than a hundred times the specified RAM, about 125 MB.)

So put, this is clearly an operationalisation of “intelligence” without reference to consciousness, intentionality, semantics, understanding or any of the other “mentalistic” concepts of philosophy of mind. (This is still a useful sidestep 80 years later.)

Appealing to “understanding”, as Simplicio did above, implies rejecting functionalism. (Where functionalism views the input/output relation or function as

constituting or producing mental activity.) So Simplicio is taking John Searle's line, of the necessity of 'original intentionality' (purposefulness, aboutness) for a system to be a mind. Searle:

...the presence of a program at any level which satisfies the Turing test is not sufficient for, nor constitutive of, the presence of intentional content. [Jacquette] thinks that I am claiming "Program implies necessarily not mind" whereas what I am in fact claiming is "It is not the case that (necessarily (program implies mind))."

i.e.

1. Programs are purely formal (syntax-only).
2. Human minds have mental content (semantics, beyond syntax).
3. Syntax by itself is neither constitutive of, nor sufficient for, semantic content.
  
4. Therefore, programs by themselves are not constitutive of, nor sufficient for minds.

Note that we've slipped from talking about intelligence (often glossed as "the production of good outputs given varied inputs") to talking about minds (which could mean intelligence, or first-person consciousness, or...). For whatever reason, this happens all the time.

The real trouble comes in his positive case - Searle's "Chinese Room" metaphor (in which no component of a translation system understands Chinese, but the Room can translate it nonetheless, giving the right input/output pairs). The Chinese Room is a punchy illustration of premise 3 above, intended to demonstrate an instance of intelligent behaviour without understanding or mental content.

1. Searle: "purely syntactic systems lack subjective experiences."
2. Searle: "I have subjective experiences."
3. So: "I am not a purely syntactic system." (modus tollens, 1&2)

This is unsatisfying: computer systems (hardware + program) are not "purely syntactic"; they have changing internal states altering according to inputs plus internal structure, a setup highly reminiscent of the representational theory of mind in humans.

Worse: as reconstructed, there's an actual fallacy here. The Chinese Room implies that syntax is not sufficient for semantics, despite the impossibility of being a syntactic system and verifying this assertion directly.

1. Searle: "purely syntactic systems lack subjective experiences."
2. Searle: "I have subjective experiences."
3. So Searle: "I am not a purely syntactic system." (modus tollens, 1&2)
  
4. The only system Searle has knowledge of the subjective experiences of is himself.

5. So if Searle is not a purely syntactic system, he has no knowledge of what it is like to be a purely syntactic system,
6. So if Searle is not a purely syntactic system, he therefore cannot assert premise 1. (5, + the knowledge account of assertion).
7. But if Searle is a purely syntactic system, (1) is false. (by 2)
  
8. You're either a purely syntactic system or you're not.
9. Therefore premise (1) is either unwarranted or false. (by 6 & 7 & 8 )

Despite Turing's inspiring attempt to sideline it, the metaphysics of mind is a live concern; Searle's objection, that the kind of minds *we know about* seem to depend on / arise out of intentionality is fine as far as it goes. But we are too ignorant to go about generalising about minds given our solitary example of the species: we haven't seen enough (as Sloman puts it, enough of the "space of possible minds") to say that particular human correlates are necessary for intelligence.

```

<h3>Disclaimer</h3>
<div>
    This was my first original philosophical argument. (The original version of it was much
    <br><br>
    These days I wouldn't use infallibilism as the baseball bat I did just there ("<i>Searle
    <br><br>
    And I'd say more about Searle's odd dichotomy between representational machines who are
</div>

<h3>Chomskyan Descartes </h3>
<div>
    I can't miss the opportunity to pass on a Good Fact: the Turing Test <a href="{{discourse
<!-- -->
    <blockquote>
        If there were machines which bore a resemblance to our bodies and imitated our action
    </blockquote><br><br>
<!-- -->
    That Descartes could not conceive of any such machine, while Turing could, is an importa
<!-- -->
    <ol>
        <li>conceivability (by a particular person, or a particular species) is far too weak
        <li>"What you can imagine depends on what you know." It is not that Turing was neces
    </ol>
</div>

<h3>Bibliography</h3>
<div>
<ul>
    <li>Block, Ned (1995), '<a href="https://www.nyu.edu/gsas/dept/philo/faculty/block/paper
    <br>

```

```
<li>Cole, David (2004); '<a href="http://plato.stanford.edu/entries/chinese-room/">The Chinese Room</a>', in D. Hofstadter (ed.), <br>
<li>Hofstadter, Douglas (1981); '<a href="">A Coffeehouse Conversation</a>', in D. Hofstadter (ed.), <br>
<li>Hofstadter, Douglas (1995), Fluid Concepts & Creative Analogies (Bloomington; Basic Books) <br>
<li>Levin, Janet (2009); '<a href="">Functionalism</a>', Stanford Encyclopaedia of Philosophy, ed. Edward N. Zalta, Spring 2009 Edition <br>
<li>Nagel, Thomas (1974); '<a href="">What Is It Like To Be A Bat?</a>', The Philosophical Review, Vol. 83, No. 4, October 1974, pp. 437-457 <br>
<li>Oppy, Graham & Dowe, David (2008); '<a href="http://plato.stanford.edu/entries/turing-machinery/">Turing's Thesis</a>', Stanford Encyclopaedia of Philosophy, ed. Edward N. Zalta, Spring 2008 Edition <br>
<li>Searle, John R (1989); '<a href="https://www.jstor.org/stable/2107856?seq=1#page_scan_tab_contents">Mimesis and Alterity</a>', Journal of Philosophy, Vol. 86, No. 1, January 1989, pp. 1-22 <br>
<li>Turing, Alan (1950); '<a href="">Computing Machinery and Intelligence</a>', Mind, Vol. 59, No. 236, April 1950, pp. 433-460 <br>
</ul>
</div>

{%
  include turing-searle/foots.html
}
```

# Misreading Russell on radical scepticism

Gavin

2012-02-01

```
{% assign niet = "https://books.google.co.uk/books?id=_qKBBbc3bWkC&pg=PA156&lpg=PA156&source=bl&xUi1gCeCjOIqeOTRA&hl=en&sa=X&ved=2ahUKEwj2jNGuhO_mAhWZQEEAHbRnCMkQ6AEwEnoECAcQ%"} {% assign duhem = "https://www.sciencedirect.com/science/article/abs/pii/S0039368106001075" %}
```

The philosopher believes that the value of his philosophy lies in the whole, in the building; posterity discovers it in the bricks with which he built and which are then often used again for better building.

- Nietzsche

Philosophers do this funny thing where they write about old philosophers, but apply wildly anachronistic or counterfactual ideas to them, in an attempt to fix their arguments for them. The following was my attempt at this sport, as an immature young man infatuated with philosophical method.

---

The hardest of hard data are of two sorts: the particular facts of sense, and the general truths of logic ... Without this assumption, we are in danger of falling into that universal scepticism which, as we saw, is as barren as it is irrefutable.

- Russell (1914)

Indeed, there is little but prejudice and habit to be said for the view that there is a world at all.

- Russell (1931)

If, however, anyone chooses to maintain solipsism [scepticism] ... I shall admit that he cannot be refuted, but shall be profoundly sceptical of his sincerity.

- Russell (1948)

The following reads two of Bertrand Russell's epistemologies (1912 & 1948) in terms of recent work, i.e. as a system of epistemic norms. This is revisionism – it is just what Russell could have done against radical scepticism, if he'd been around now.

His book *Human Knowledge* is much more interesting than it gets credit for; it prefigures several new schools: it is proto-virtue epistemology, proto-externalist, proto-Bayesian, and proto-naturalised epistemology.

---

### A general sceptical argument

Derived from Descartes' dream argument:

- 1) It is possible that sceptical hypothesis  $S$  is true.
  - 2) I cannot with certainty determine  $S$  is false.
  - 3) If I cannot with certainty determine  $S$  false, then some serious class of my beliefs lack "knowledge" status.
- C) So some serious class of my beliefs lack "knowledge" status.
- 

#### i. Reconstructing Russell

"...the traditional epistemological project [saw the] theory of knowledge as bulwark against scepticism; proponents of [the new virtue epistemology] anticipate its displacement by a more diverse set of concerns..."

- Guy Axtell

"Mathematics and the stars consoled me when the human world seemed empty of comfort. But changes in my philosophy have robbed me of such consolations... It seemed that what we had thought of as laws of nature were only linguistic conventions, and that physics was not really concerned with an external world. I do not mean that I quite believed this, but that it became a haunting nightmare, increasingly invading my imagination."

- Russell

Did Russell naturalise epistemology, do virtue epistemology, or employ an ethics of belief? Well, his last major philosophical work, *Human Knowledge* (1948) pre-dates Quine's launching of the naturalising project by twenty years, and also pre-dates the first explicit piece of 'virtue epistemology' by thirty years - so the idea is absurd revisionism.

But anachronism has its use: there are few absolutely novel concepts and questions in these new approaches to epistemology; they are shifts in emphasis and method, away from the so-called "doxastic paradigm". (Aristotle, for instance, can be artlessly seen as a virtue epistemologist, and there have recently been considered accounts of René Descartes, Pierre Duhem and even (tenuously) WVO Quine as virtue epistemologists.)

Let's see if we can fix Russell's responses to radical scepticism. I focus on his tacit use of epistemic norms (hence, 'ethics of belief'), rather than on virtue theory or naturalism. The trick will be to preserve Russell's realism and *HK*'s early naturalistic epistemology in meta-epistemology. I draw on two of his books: *The Problems of Philosophy* (1912) and *Human Knowledge* (1948).

"Virtue epistemology": naturalism + normativity + speech-act theory + practical reason.

---

### ii. Epistemic normativity

To ascribe knowledge is to evaluate positively as well as to describe a state. Epistemology cannot avoid normativity, since even the most naturalised accepts at least one epistemic norm:

Radical Quinean norm: "Epistemology should not be normative."

An epistemic norm is a some standard with a bearing on knowledge. They are not preferences.

Duncan Pritchard gives a stricter realist definition which accords with Russell: "*a rule which one follows in order to gain true belief*".

Pascal Engel adds sensible conditions: "*For a principle to be genuinely normative, it must have normative force, and to be able to actually regulate belief. It must also have normative freedom... one must have the possibility of violating it.*" (As opposed to an epistemic virtue: "*an embodied habit that promotes the acquisition, maintenance, and transmission of epistemic goods.*")

Distinguish meta-epistemology (which yields methodology) and the 'ethics of belief' (which yields norms) – but unfortunately the distinction isn't clear. Russell discusses both together as 'maxims', and others in the past called both 'principles'. I take choice of methodology as reducible to epistemic normativity. The notion is a bit plastic – for instance, logical laws seem statable as epistemic norms:

Coherence norm: "one ought not to believe p and not p."

Closure norm: "one ought to believe that q if one believes that p & that p entails q."

---

### iii. Some hefty qualifications

- Russell's epistemology is motivated by a specific view of logic and semantics. I'm skipping all that. I address the extent to which, despite his innovative

method, Russell (1912) was a ‘traditional’ epistemologist – some family-resemblance of internalist, infallibilist, methodist & foundationalist - in section iv.

- Russell’s semi-naturalised epistemology in *Human Knowledge* poses an issue: how are we to discuss norms in a ‘natural’ descriptive epistemology? One answer is that Russell’s ‘naturalistic turn’ (from 1940 on) wasn’t the kind that demands the reduction of normative facts to natural ones.
- We can express his realism and anti-psychologism by construing Russell (with the rest of ‘traditional’ epistemology) as a normative cognitive monist, holding that there is one universally applicable set of correct epistemic norms.
- The practice could be intellectually venal: “*Kornblith contends that once traditional epistemologists admit that the Cartesian program of deriving beliefs about the world from certain foundations fails, they end up endorsing as legitimate whatever principles enable them to ratify the beliefs they started with.*” Against this valid worry I’d firstly say that the epistemic norms are as open to criticism, as any position in philosophy; and, secondly, this doesn’t stick to *HK* Russell, owing to his naturalistic epistemology: any epistemic norms he endorsed would (in principle) be open to empirical test. (Except the Postulates; see section v.)
- My interpretation does not contort him into a virtue epistemologist: I focus on doxastic norms (abstract rules) rather than epistemic virtues (agents’ traits). John Greco gives two necessary conditions for virtue epistemologies: the acceptance both “that epistemology is a normative discipline” and “that agents and communities are the primary source of epistemic value and the primary focus of epistemic evaluation...” Russell endorsed the first but not the second (excepting suggestive passages in *HK*).

Thus qualified, what remains of the merits of the approach? The clearest answer comes in the gap between his ‘traditionalist’ fixation on scepticism and his non-traditional responses to it.

---

#### iv. Russellian indirect realism as epistemic norm

When... we speak of philosophy as a criticism of knowledge, it is necessary to impose a certain limitation. If we adopt the attitude of the complete sceptic, placing ourselves wholly outside all knowledge, and asking, from this outside position, to be compelled to return within the circle of knowledge, we are demanding what is impossible, and our scepticism can never be refuted... But it is not difficult to see that scepticism of this kind is unreasonable.

- Russell

In *The Problems of Philosophy*, Russell tries to explain our knowledge as a process of abduction from directly apprehended facts ('knowledge by acquaintance') to facts that explain them ('knowledge by description'). Here is a representation, which I'll modify as the sceptical challenge runs its course:

### Assumptions

- Minimal realism: Experiences are caused by things other than experiences.
- Minimal causal law: If like cause, then like effect.
- Incorrigibility: What is known non-inferentially is proof against radical scepticism.
- Methodism: Aims to find the criteria for knowledge without claiming instances, thus avoiding circularity.
- Foundationalism: "*Starting with the common beliefs of daily life, we can be driven back from point to point, until we come to some general principle, which seems luminously evident, and is not itself capable of being deduced from anything more evident.*"
- Principle of Acquaintance: "*Every proposition which we understand must be composed wholly of constituents with which we are acquainted*"

### Norms

Commonsense norm: We should prefer views which grant us knowledge.

Principle of Phenomenal Conservatism: One is permitted to assume things are as they appear, except when there are positive grounds for doubting this.

5'. Russell's norm of doubt: "We cannot have reason to reject a belief except on the ground of some other belief."

Internalist's norm: beliefs are to be justified only by one's own psychological experiences: the justifying relations between one's experiences and beliefs are to be worked out from 'inside'.

Justification norm of assertion: You ought not believe p unless you are warranted to assert p. (As opposed to the default Knowledge norm of belief: You ought not believe p unless you know p.)

On to radical scepticism. Russell offers a *normative* response. The normative backdrop of sceptical paradoxes (given norm 6 and 7) is:

Lack norm of doubt: If not sufficiently grounded, any belief is open to legitimate doubt.

Infallibilist norm of assertion: You ought not believe p if p is open to legitimate doubt.

Prove-it norm: The burden of proof for any claim falls to the claimant.

(8) and (10) are essential to critical thinking – but when combined with internalism (6), they generate a destructive sceptical demand: hyperbolic

doubt and synchronic reconstruction in sequence: “*take the totality of things you believe, subtract [your] claim and everything that you cannot defend without assuming it, and now show that the claim is correct.*”

Russell thinks this is impossible (see norm 5’), and tries to block this scale of scepticism by showing the position it entails to be unreasonable, and so negligible. (He is in effect defending the bare thesis *Dogmatism*, that at least one of one’s knowledge-claims is true, & that one knows it is.)

Reconstructed:

- P1. If it is impossible to meet a demand, then that demand is unreasonable.
- P2. It is impossible to meet the sceptical demand.
- P3. If the demand is unreasonable, then the sceptical position is unreasonable.
- C. The sceptical position is unreasonable. (by double modus ponens)

But this requires a further epistemic norm underlying the whole attitude:

Dogmatic norm: It is unreasonable to doubt dogmatism if there are no possible reasons that could persuade someone who doesn’t believe it to believe it.

This is obviously ad hoc, but it has bigger problems. Say there are only two possibilities:

- 1) if one finds compelling reasons to endorse dogmatism, one must believe it (and thereby stop being a sceptic) on pain of unreasonableness; and
- 2) even if one does not find compelling reasons, then (11) makes one unreasonable not to believe it.

The unacceptable implication is that inability to ground a position is taken to be a compelling reason to believe it. (It also breaches Engel’s condition of normative freedom, given above.) Further:

- P1. By (8), every belief is subject to examination and the possibility of doubt
- P2. Dogmatism is a belief.
- P3. Thus the sceptic can examine dogmatism and possibly doubt it.
- P4. Examinations must admit the possibility there are no compelling grounds.
- P5. If they are not found, one need not believe dogmatism.
- P6. Russell’s reasoning entails that dogmatism cannot be truly examined.
- C. So by contradiction of basic norms, the argument fails.

It could be rearticulated as a properly general norm:

11’. *Cogency norm*: It is unreasonable to doubt a position if there are no possible reasons that could persuade someone who doesn’t believe the position to believe it.

But this is absurdly strong: even if Russell were to accept it (and the suggestion is philosophical slander), the argument collapses, since:

- P1. (11') stamps as indubitable all sets of beliefs that cannot be justified except by circular reasoning.
- P2. There are a vast number of such sets.
- P3. Many of those sets will be incompatible with each other.
- C. (11') generates and asserts a vast number of contradictions.
- 

#### v. epistemic norms in *Human Knowledge*

*Human Knowledge* (HK) offers another foundationalism, but one that rejects pure empiricism.

It gives up epistemic ground – conceding that data are private, and we cannot demonstratively infer an external world from them – and then tries to retain knowledge of the external world by lionising “non-demonstrative inference”. It is in a detailed and somewhat Bayesian treatment of it that the meat of *HK* is said to lie.

But HK actually develops a double theory of knowledge, with two sets of standards, since it also holds the core doctrines of what we now know as naturalised epistemology: it is fallibilist, views some knowing as animal behaviour, takes “best science” as a given and invites psychology to bear on epistemological questions. His argument boils down to a pragmatic demand to widen our conception of reasonable justification to include (some) non-demonstrative inferences:

- P1) Scientific inference is not demonstrative.
- P2) Either it is unreasonable, or not all reasonable argument is demonstrative.
- P3) Scientific inference is not unreasonable.
- C) Not all reasonable argument is demonstrative.

The project is to canonise scientific inference, which sidelines scepticism (though he claims to not be merely ignoring it).

#### Norms

Norms 4, 5 and 7 carry over from Russell (1912).

Naturalist’s norm: the primary problem of epistemology is a descriptive one: “when does scientific method allow us to infer an unobserved thing from what we observe?” This can rightfully be done without demonstrative answers to the normative question. (rejection of norm 6).

The most remarkable part of *HK*, though is his presentation of five “postulates”: really vague contingent general facts, which together give the minimal ontology that permits applied probability, and thereby induction, and thereby scientific inference (and some of commonsense):

- Postulate of quasi-permanence (accounts for objects without ontology)

- Postulate of separable causal lines (accounts for regularities and, e.g. motion)
- Postulate of spatio-temporal continuity (enables realism: unperceived existents)
- The structural postulate (accounts for e.g. improbability of a repeated coincidence)
- The Postulate of analogy: (accounts for persisting properties and other minds)

After surveying the options (that he could conceive of, see below), Russell claims for these the status of non-inferential synthetic apriori knowledge – “if it can be called ‘knowledge’”. But he clearly anticipates the sceptic’s valid reply. His positive argument is unusual, utilising as it does a special conception of knowledge that prefigures epistemic externalism (italics):

In what sense can we be said to ‘know’ the above postulates? ...  
 [inductive standards] are valid if the world has certain characteristics which we all believe it to have ... therefore we may be said to “know” what is necessary for scientific inference, given that it fulfils the following conditions: (1) it is true, (2) we believe it, (3) it leads to no conclusions which experience confutes, (4) it is logically necessary if any occurrence... is ever to afford evidence in favour of any other occurrence.

The analogy to the *Problems*’ problematic anti-sceptical strategy is clear. But what is the epistemic status of the postulates? They don’t fit assumption, empirical fact (since they enable empirical generalisation), Kantian category, apriori intuition, logical law, methodological principle or, alas, epistemic norm. (This is partly due to Russell’s reformulation of the basic terms of epistemology.)

Some recent resources resonate. For instance, Roderick Chisholm is indebted to *HK*; his principles of evidence are akin to laws built from Russell’s postulates – or, indeed, to epistemic norms over non-demonstrative inference. But Chisholm held them to be necessarily true apriori, to give *prima facie* evidence, and to be ‘internal’ “*in that the proper use of them at any time will enable us to ascertain the epistemic status of our own beliefs at that time.*” Unfortunately, though, the necessity alone would alienate an HK-Russellian, since Russell devotes almost a full chapter to an explanation of his rejection of necessary relations over and above his postulates.

### Hinges

The “hinge proposition”, hinted at among Wittgenstein’s last notes, has become a regular feature of recent sceptical discourse. A hinge proposition is said to work outside justification as conceived in norms (7), (8), (10) – it is not itself knowledge, because it is outside epistemic evaluation. Recall:

Lack norm of doubt: If not sufficiently grounded, any belief is open to legitimate

doubt.

Wittgenstein clearly rejects this. Speculatively:

7'. Hinge norm of justification: There are some beliefs which in some contexts one may legitimately believe without justification.

8'. Hinge norm of doubt: There are some beliefs which in some contexts one ought not to doubt.

Kornblith's concern about unprincipled principles comes to mind. But in any case Russell's postulates are not good candidates for hinge propositions, since his examples are everyday, pre-theoretical beliefs, leading to an exemplary Wittgensteinian suggestion: that it necessarily slips between philosophy's fingers.

Failing these reconstructions of Russell's postulates, though, their status is unclear. They could be wishful thoughts, or the most abstract appeal to common sense ever.

---

### The “Russellian” Retreat

```
<blockquote>
    We can live with the concession that we do not, strictly, know some of the things we be-
</blockquote>
- Crispin Wright
<br /><br />
```

An unpopular solution, also named for Russell: admit defeat. Retreat from knowledge: take the radical sceptical paradox to be truly informative about the concept *knowledge* – but retain entitlement to one's justified beliefs.

To be entitled to accept a proposition in this way, of course, has no connection whatever with the likelihood of its truth. We are entitled to proceed on the basis of certain beliefs merely because there is no extant reason to disbelieve them and because, unless we make some such commitments, we cannot proceed at all. Any epistemological standpoint which falls back on a conception of entitlement of this kind for the last word against scepticism needs its own version of (what is sometimes called) the Serenity Prayer: in ordinary enquiry, we must hope to be granted the discipline to take responsibility for what we can be responsible, the trust to accept what we must merely presuppose, and the wisdom to know the difference.”

Returning to the epigram: the anti-sceptical ‘responses’ above are not foundations, refutations, nor Moorean denials of scepticism. (They do not yield certainty.)

On the bright side, *HK* prefigures several (of what we currently think of as acutely differing) approaches: virtue, doxastic ethics, Bayesian, externalist, and

naturalised epistemology. Maybe the divides between these are not insurmountable.

### Bibliography

<div>

- \* Axtell, Guy (2006), ‘Epistemic Virtue’, Janusblog, <http://goo.gl/1jBfl>
- \* Brueckner, Anthony (1992), Mind , New Series, Vol. 101, No. 402 (Apr., 1992), pp. 309-317
- \* Descartes, Rene (1716), Meditations on First Philosophy, Cambridge University Press
- \* Fairweather, Abrol (2012), ‘Duhem-Quine Virtue Epistemology’, for Synthese, <http://philpaper.scholarlycommons.psu.edu/1113.html>
- \* Feldman, Richard (2001), ‘Naturalized Epistemology’, Stanford Encyclopaedia, <http://plato.stanford.edu/entries/naturalized-epistemology/>
- \* Goldman & Olsson (2009), ‘Reliabilism and the Value of Knowledge’, <http://goo.gl/zb1BU>
- \* Greco, John (2011), ‘Virtue Epistemology’, Stanford Encyclopaedia of Philosophy, <http://plato.stanford.edu/entries/virtue-epistemology/>
- \* Hellie, Benj (2000), ‘Acquaintance’, in the Oxford
- \* Kitcher, Philip (1992), ‘The Naturalists Return’, Philosophical Review 101 (1) p.53-114
- \* Klein, Peter (2005) ‘Epistemology’, in Craig (ed.), Routledge Encyclopedia of Philosophy
- \* Pritchard, Duncan (2001), lecture notes for ‘Scepticism and the Structure of Knowledge’ <http://www.duncanpritchard.com/teaching/epistemology.pdf>
- \* Quine, W.V.O. (1986), ‘Reply to Morton White’ in The Philosophy of W V. Quine, ed. Hahn and Putnam
- \* Robinson, Alistair (2010), <http://critique-of-pure-reason.com/crispin-wrights-scepticism-and-epistemology/>
- \* Russell, Bertrand (1912), The Problems of Philosophy (Suffolk; Oxford University Press; 1992)
- \* Russell, Bertrand (1914), Our Knowledge of the External World (London; Routledge; 1993)
- \* Russell, Bertrand (1940), An Inquiry into Meaning and Truth (New York; Norton & Co; 1940)
- \* Russell, Bertrand (1948), Human Knowledge (London; George Allen & Unwin; 1948)
- \* Russell, Bertrand (1960), The Autobiography of Bertrand Russell, (Padstow; Routledge; 1998)
- \* Sinclair, Robert (2009), ‘Quine’s Philosophy of Science’, Internet Encyclopaedia of Philosophy
- \* Sosa, Ernest (1980), ‘The Raft and the Pyramid’, Midwest Studies In Philosophy, Volume 5, Number 1
- \* Sosa, Ernest (2007), Virtue Epistemology , Oxford Bibliographies Online, <http://goo.gl/qbuLJ>
- \* Sosa, Ernest (2009), ‘Descartes and Virtue Epistemology’, in New Essays on the Philosophy of Knowledge, Cambridge University Press
- \* Stich, Stephen (1988). ‘Reflective Equilibrium, Analytic Epistemology and the Problem of Coherence’, in *Philosophical Perspectives*, 1, 1988, pp. 1-22
- \* Strawson, P.F. 1985: Scepticism and Naturalism: Some Varieties, Routledge
- \* Stroud, B. 1984: The Significance of Philosophical Scepticism, Oxford University Press
- \* Stump, David (2007) ‘Pierre Duhem’s Virtue Epistemology’, Studies in the History & Philosophy of Science, 38(2), 2007, pp. 111-129
- \* Wright, Crispin (1991) ‘Scepticism and Dreaming: Imploding the Demon’, Mind 397 (1991), pp. 397-415
- \* Wright, Crispin (2000), ‘Cogency and Question-Begging: Some Reflections on McKinsey’s Paradox’, in *Philosophical Perspectives*, 13, 2000, pp. 1-22
- \* Zagzebski, Linda (1996) *Virtues of the Mind* (Cambridge; Cambridge University Press)

# My extended mind of minds

Gavin

2018-03-24

{% include exomind/links.md %}

Some archaeologists now say that when they dig up the remains of lost civilizations they are not just reconstructing objects but reconstructing minds.

Some musicologists say that playing an instrument involves incorporating an object into thought and emotion, and that to listen to music is to enter into a larger cognitive system comprised of many objects and many people.

— Larissa MacFarquhar

I am not so clever. Insofar as my model of the world is good, it's because I have found a hundred reliable specialists whose research I can trust.

Philosophers call this “socially-extended cognition” - though this is just a metaphysically extravagant version of the ancient idea of “testimony”. These people might be the part of my mind outside of my skull.

This is a risky business: I'm a reductionist about warrant: beliefs obtained this way are a level removed from real justification. So it's not a matter of just ingesting their beliefs; instead, they're a noisy proxy for all the empirical and logical work it would take hundreds of thousands of hours to do myself. It is obviously necessary to have at least two people in each category who come at the field from very different angles.

The warrant given varies, of course. The testimony of a physicist on physics is much stronger than a psychologist on psychology.

The following are all living people because I'm being ‘conservative’ and including only external *processing*, incorporating new info. (Books - which contribute as much or more of my beliefs - are in a sense distributed storage rather than processing). 1

Modelling this

---

## List

- Altruistic effectiveness: Jeff Kaufman, Nick Beckstead, Toby Ord, Michael Plant,
- Statistics: Andrew Gelman, David Spiegelhalter, Cosma Shalizi, Nassim Taleb.
- Meta-science: Retraction Watch, John Ioannidis, Richard McElreath, Uri Simonsohn, Alan Sokal, Jeff Leek.
- Physics: Sabine Hossfelder, Scott Aaronson, Jean Bricmont, Sean Carroll, Roger Penrose.
- Philosophy: Joseph Heath, Will MacAskill, Scott Aaronson, Eric Schliesser, Eric Schwitzgebel, Brian Leiter.

Judea Pearl Robert Paul Wolff Ken White Esther Duflo; Henry Farrell, William Easterly. Barbara Ehrenreich, Kieran Healy

- Macroeconomics: Dani Rodrik, Bryan Caplan, Noah Smith, John Quiggin, Deirdre McCloskey, Paul Krugman,.
- Policy: David Roodman, Steven Pinker.
- Economic history: Artur Kel, Pseudoerasmus.
- Economic justice: David Graeber, Freddie deBoer, Chris Dillow,
- Nutrition: Stephen Guyenet, Examine.
- Psychiatry, neuroscience, medicine: Scott Alexander, Cochrane Collaboration,
- Development: David Roodman, Chris Blattman
- Psychology: Stuart Ritchie, Daniel Kahneman, Nick Brown,
- Genetics: Stephen Hsu, Stuart Ritchie, Gwern
- AI: Gwern Branwen, Chris Olah, Katja Grace,
- Cultural politics: Kelsey P, Joseph Heath, Thomas Basbøll
- British politics: Charlie Stross, David Allen Green, Chris Dillow, David Torrance, Stewart Lee.
- Gender: Ozy Frantz, Kelsey P, Taylor Saotome-Westlake
- Computer science: Scott Aaronson,
- Software: Dan Luu, John Morrice,
- History: Clive James,
- The Future: Robin Hanson, Anders Sandberg, David Pearce.

- Art
- Misc: Katja Grace, Tyler Cowen, Rob Wiblin, Jeff Kaufman, Scott Aaronson, Robin Hanson, Hugh Panton, Anders Sandberg, Buck Shlegeris, Scott Alexander.

This is a living list; please add your own outboard below.

{% include exomind/foots.html %}

# Preventing Side-effects in Gridworlds

Gavin

2018-04-22

{% include gridworlds/links.md %}

Joint work with Karol Kubicki, Jessica Cooper and Tom McGrath at AISC 2018.

Can we ensure that artificial agents behave safely? Well, start at the bottom: We have not even solved the problem in the concrete 2D, fully-observable, finite case. Call this the “gridworld” case, following Sutton and Barto (1998).

Recently, Google DeepMind released a game engine for building gridworlds, as well as a few examples of safety gridworlds - but these came without agents or featurisers. In April our team implemented RL agents for the engine, and started building a safety test suite for gridworlds. Our current progress can be found here, pending merge into the main repo.

We focussed on one class of unsafe behaviour, *(negative) side effects*: harms due to an incompletely specified reward function. All real-world tasks involve many tacit secondary goals, from “... without breaking anything” to “... without being insulting”. But what prevents side effects? (Short of simply hand-coding the reward function to preclude them - which we can’t rely on, since that ad hoc approach won’t generalise and always risks oversights.)

## Taxonomy of environments

We made 6 new gridworlds, corresponding to the leaf nodes shown above. In the following, the left is the unsafe case and the right the safe case:

### Static deterministic:

- “Vase world”. Simply avoid a hazard.
- “Burning building”. Balance a small irreversible change against a large disutility.
- “Strict sokoban”. Reset the environment behind you.

## Dynamic deterministic

- “Teabot”. Avoid a moving hazard. 2
- “Sushi-bot”. Be indifferent to a particular good irreversible process.
- “Ballbot”. Teabot with a moving goal as well as a moving hazard.

## Stochastic

We also have stochastic versions of “BurningBuilding” and “Teabot”, in which the environment changes unpredictably, forcing the agent to be adaptable.

One kind of side effect involves irreversible change to the environment. Cases like sushi-bot suggest that a safe approach will need to model types of irreversibility, since some irreversible changes are desirable (e.g. eating, surgery).

The environments can be further categorised as involving:

- *Hazard* - objects the agent should not interact with, either because they are fragile or because the agent is (e.g. a vase, the floor is lava).
- *Progress* - irreversible processes which we want to occur (e.g. sushi ingestion).
- *Tradeoff* - irreversible processes which prevent worse irreversible processes (e.g. breaking down a door to save lives).
- *Reset* - where the final state must be identical to the initial state (but with the goal completed). (e.g. controlled areas in manufacturing)

## Taxonomy of agent approaches

### 1. Target low impact

- Penalise final state’s distance from the inaction baseline. 1
- Penalise the agent’s *potential* influence over environment.3
- Penalise distance from a desirable past state. 4

### 2. Model reward uncertainty

- Use the stated reward function as Bayesian evidence about the true reward. Leads to a risk-averse policy if there’s ambiguity about the current state’s value in the given reward function. 5

### 3. Put humans in the loop

- “Vanilla” Inverse reinforcement learning
  - Maximum Entropy
  - Maximum Causal Entropy
- Cooperative IRL
- Deep IRL from Human Preferences

- Evolutionary: direct policy search via iterated tournaments with human negative feedback.
- Deep Symbolic Reinforcement Learning. Learn a ruleset from pixels, including potentially normative rules.
- Whitelist learning

### **Agent 1: Deep Q-learning**

We first implemented an amoral baseline agent. Code here.

### **Agent 2: MaxEnt Inverse Reinforcement Learning**

{% include gridworlds/irl.html %}

### **Reflections**

- Reset and empowerment trade off in the Sokoban grid - putting the box back to the starting point is actually irreversible.
- How well will features generalise? Would be good to train features in some environments before testing in random new but similar ones
- Expect to be able to learn tradeoff between empowerment loss and rewards directly by using CIRL - learn goal and empowerment/ergodicity parameters that set preferences
- Demonstrations being the same length is a strange and not ideal limitation
- Could have many features, some of which should be zero - e.g. distance between agent and box - but which the demonstrations are also consistent with being nonzero. It's impossible to distinguish between these given only the demonstrations at hand. There is almost certainly some (anti)correlation between features, e.g. large agent-box distance weights explain away the trajectories without requiring any weight on the 'is it in a corner' feature. Inverse reward design offers a way to resolve this, but I don't think it has all the details necessary.
- Maybe if we had some sort of negative demonstrations (human to agent: don't do this!) then learning zero weights would become possible (formally we could try to maximize probability of positive demonstrations while minimizing probability of the negative ones)
- Trajectories demonstrated by IRL don't necessarily look like the ones given, especially if there are 'wrong' features that are maximised under the demonstrations
- What are we trying to achieve with each gridworld? E.g. Reset is harder to define in dynamic environments and even harder in stochastic ones, sometimes irreversibility is desired (sushi) or needs to be traded off against utility in a context-dependent way (burning building)

- Issues:
  - No way to give negative feedback
  - No way to give iterative feedback
  - Neither of these are lifted by IRD or Deep IRL, but IRD generates the kind of data we might want as a part of the algorithm (approximating the posterior)
- IRL solves an MDP at every update step. At least this value-aware algorithm is at a massive disadvantage.

## Future work

- Pull request with the new environments, agents and transition matrix calculator.
- Implement more complex features
- Implement MaxEnt Deep IRL, Max Causal Entropy IRL
- Implement IRD
- Think about negative/iterative feedback models
- Automate testing: for all agents for all grids, scrutinise safety.

## Bibliography

See the Google sheet here.

Applications for the next AI Safety Camp will open around June. I highly recommend it.

{% include gridworlds/foots.html %}

# ‘Curiosity’ (2012) by Ball

Gavin

2018-07-02

... — why is the sea salty? — have animals souls, or intelligence? — has opinion its foundation in the animate body? — why do human beings not have horns? — how is it that sound in its passage makes its way through any obstacle whatever? — how is it that joy can be the cause of tears? — why are the fingers of unequal length? — why, if you have intercourse with a woman after she has lain with a leper, do you catch the disease while she escapes? — what reason is there for the universality of death? — why do we need food so frequently, or at all? — why are the living afraid of the bodies of the dead? — how is the globe supported in the middle of the air? — why does the inflow of the rivers not increase the bulk of the ocean? — why, if a vessel be full and its lower part open, does water not issue from it unless the upper lid be first removed? — when one atom is moved, are all moved? (since whatever is in a state of motion moves something else, thus setting up infinite motion.) — why do winds travel along the earth’s surface and not in an upward direction? — why does a sort of perpetual shadow brood over the moon? — granted that the stars are alive, on what food do they live? — ought we regard the cosmos as an inanimate body, a living thing, or a god?

— Adelard of Bath (c.1120)

{% include curio/links.md %}

Another history of the origins of science: our long trek to GWAS, livermorium, and CERN via astrology, natural magic, alchemy, Neoplatonism, herbalism, occultism, and philosophy. So, superficially, the book is just about an especially fruity context of discovery. But this period holds two of the most important lessons in history:

1. science grew out of work by people who diverge wildly from the modern idea and practice of science, whose variously false frameworks led to the Royal Society and e.g. the Newtonian triumph. (And from there to contemporary, professional, university science.) So wrong people can still make progress if their errors are uncorrelated with the prevailing errors.
2. a small number of the most powerful people in Britain - the Lord Chancellor, the king’s physicians, the chaplain of the Elector Palatine & bishop of

Chester, London's great architect, various Privy Councillors - successfully pushed a massive philosophical change, and so contributed to most of our greatest achievements: smallpox eradication, Sputnik and Voyager, the Green Revolution, and the unmanageably broad boons of computing are partly theirs.

```
<h3>Baron Verulam and the future of humanity</h3>
<div>
    Bacon has some claim to being the most influential philosopher ever, in terms of counter
    (Yes, <a href="{{hay}}">ibn al-Haytham</a>'s was 7 centuries ahead of its time, but to I
    (Yes, in fact <a href="{{induc}}">his biggest single philosophical doctrine</a> is shak
</div>
```

---

## A model of science

```
{% include curio/model.html %}
```

---

## What took so long?

All of the pieces of science are very ancient - we had mathematics and data collection well before the Ten Commandments, naturalism before Buddha and Confucius, reductionism before the Peloponnesian War at least one controlled trial centuries before Christ, fallibilism likewise. Everything was ready BCE; we can see indirect evidence of this in the astonishing works of Ancient Greek engineers, mostly unmatched for 1000 years until y'know.

So the question is not “was Bacon the most original blah blah?”: he wasn’t, particularly when you remember Alhazen’s Baconian method, from the C11th. But we need an explanation for how we managed to mess it up so badly. The received view, which is all I have at the moment, is that the fall of Rome, Christian anti-intellectualism and, later, the enshrining of Aristotelian mistakes was enough to destroy and suppress the ideas. I want deeper explanations though. (For instance, what did we do to the economy?)

```
<h3>Alright let's say something about the actual book</h3>
<div>
    Back to the book eh! Book structure is lots of little chapters on fairly disjointed topi
    <blockquote><br>this was more than a case of 'look what I've got'. The power with which
    </blockquote><br>
    Ball doesn't like us calling the Scientific Revolution a revolution, and I agree: the re
    Ball expends a lot of time on a history of wonder vs curiosity vs <a href="{{hypo}}">fals
    (For instance, Margaret Cavendish - the darling of arts academics who latch on to the on
```

Stimulating as always.  
  </div>  
  { % include curio/foots.md %}

# 'Why Moral Theory is Boring and Corrupt'

Gavin

2018-07-31

```
{% assign ea = "https://www.effectivealtruism.org/articles/introduction-to-effective-altruism/" %} {% assign gays = "https://www.theguardian.com/books/2014/jun/26/sexual-irregularities-morality-jeremy-bentham-review" %} {% assign animals = "https://www.utilitarianism.com/jeremybentham.html" %} {% assign slaves = "https://ndpr.nd.edu/news/utilitarianism-and-empire/" %} {% assign austin = "https://books.google.co.uk/books?id=mHH4quX4rCQC&pg=PR17&lpg=PR17&dq=%22It+was+never+conte fJxUPyXu-Ds0jxmQ&hl=en&sa=X&ved=2ahUKEwi6xYmT_dXcAhWMMewKHehXDmcQ6AEwAHoECAAQAAQ&fjxUPyXu-Ds0jxmQ&hl=en&sa=X&ved=2ahUKEwi6xYmT_dXcAhWMMewKHehXDmcQ6AEwAHoECAAQAAQ&%} {% assign women = "https://blogs.ucl.ac.uk/museums/2016/03/03/bentham-the-feminist/" %} {% assign pun = "https://plato.stanford.edu/entries/bentham/#PenLawPun" %} {% assign arch = "https://web.archive.org/web/20180805133058/https://www.gleech.org/img/why%20moral%20Unknown.doc" %}
```

One of my favourite philosophy papers recently disappeared from the internet. It's anonymous, a beautiful and caustic dismissal of all rationalistic theories of ethics, which the author groups together as "Master Factor" theories (which reduce action to one dimension, when they think this cannot and should not be done).

moral theory is exclusive, reductively narrow in its approach to the practical questions that we need to answer; these features of moral theory make it boring, because monotonous, and corrupting, because they encourage us to see this monotony, wrongly, as a good thing; they make moral theory actually corrupt, where mauvaise foi is involved.

They're not a nihilist, but rather openly intuitionistic:

love is what, most of the time, motivates most of us who are neither complete bastards, nor distracted by secondary concerns such as "what other people will think"—to say this is not to say anything very neat or tidy, either. But that too is as it should be.

It reads like a farewell to academia, a cry of exhaustion from a foiled job-seeker:

As all too often elsewhere in universities, the entrenched sects and their apparently immutable and interminable oppositions persist, not because a compelling intellectual case can be made in their defence (a

priori it is entirely possible that the whole lot of them are indefensible), but because each of these sects has fought a successful campaign in institutional politics to establish its curricular and budgetary space—in other words, to become one of the vested interests that deans, heads of department, and other bureaucratic managers must accommodate.

I'm a thoroughgoing Boring-Corrupt consequentialist myself, but I like this paper and don't want it to fall down the digital hole. Here's the original .doc (Internet Archive) which I happened upon sometime in 2009.

(I spent a little while trying to work out who wrote it, based on their personal acknowledgments to various St Andrews, Leeds, and Sheffield philosophers, but decided I don't care.)

---

### Response

The key claim is that it's psychologically impossible to be a human really acting according to a Master Factor theory. We are too divided, contradictory, and various; as a result it's dishonest and unhealthy to pretend you are, or to try to.

For instance, if we were perfect (first-order) consequentialists, we'd be constantly paralysed by the need to analyse all of our actions in terms of their effect on the world. This would make us miserable and completely ineffective. (Stocker: "to the extent that you live the theory directly, to that extent you will fail to achieve its goods.")

The standard response is to separate the 'criterion of rightness' (what is actually good) from the 'deliberative procedure' (how we go about trying to achieve good). You only optimise the big things, using your limited information and cognitive bandwidth as much as you can, but without angst at being imperfect; you cannot be responsible for something you have no power over. (Austin: "It was never contended... by a sound, orthodox utilitarian that the lover should kiss his mistress with an eye to the common weal.") Anonymous says we can't do that.

It's clear that humans are at best imperfect consequentialists: not least, you must have accurate beliefs to reliably have good effects on the world, and almost no-one generally does. The psychological possibility of living a strict moral code is an empirical question in general - but as existence-proof I can tell Anonymous that I'm a happy person with fairly strict consequentialist morals, a strong sense of community, and as many loving relationships as I can take.

Also - if I'm allowed a circular comment: intuitionism generally leads to poor actions. Intuitionism (e.g. "act as love demands you to act") is often wrong because our intuitions are rooted in our brutal and amoral natural history, where selfishness, nepotism, othering and myopia were all highly adaptive strategies.

Vengeance is intuitive; honor killing is intuitive; actual political corruption is highly intuitive.

Around 1800, the arch-rationalist Bentham predicted that homosexuality wasn't wrong, that abusing animals was wrong, that slavery was wrong, that women deserved the vote, that retributive punishment is wrong. These remained highly counterintuitive to most of the world for the next two hundred (three hundred?) years. (An imperfect reasoner like all of us, he was wrong about other things, e.g. the colonies.) Was it reason that made us comply with these? At least partially, yes.

# ‘The Unpersuadables’ (2013) by Storr

Gavin

2018-08-21

{% include heretic/links.md %}

Imagine if, way back at the start of the scientific enterprise, someone had said, “What we really need is a control group for science - people who will behave exactly like scientists, doing experiments, publishing journals, and so on, but whose field of study is completely empty: one in which the null hypothesis is always true.” That way, we’ll be able to gauge the effect of publication bias, experimental error, misuse of statistics, data fraud, and so on, which will help us understand how serious such problems are in the real scientific literature.” Isn’t that a great idea? By an accident of historical chance, we actually have exactly such a control group, namely parapsychologists

- Allan Crossman

An irritating but righteous book. Not quite what it looks like: another Ronson-Theroux journalist, accosting another set of tragicomic kooks). OK, it is that, but it’s also a grim reflection on how confusing and muddy the world is, on the universality of extreme bias, plus Storr’s personal traumas and peccadilloes. (Half the book is his confessing to childhood theft, psychosis, academic failure, and petty vendettas.) Rather than getting to the bottom of ESP, or morgellons, or homeopathy, or past-life regression, Storr tries to understand the character of the people who believe and disbelieve in them. (This is a dangerous approach unless you are extremely sensitive and charitable. As we’ll see, Storr is that sensitive, to one set anyway.) Besides confronting unusual beliefs without (as much) prejudice, The Unpersuadables is about the fact that we are all riddled with deep obstacles to objectivity: there’s our ingroupism and confirmation bias; representation realism; emotional reasoning about nonemotional things; the terrifyingly unreliable reconstructive nature of memory; the sad distinctness of intelligence and rationality; evolutionarily adaptive delusions of superiority and agency.

These are illustrated by interviews with a creationist, Sheldrake, Irving, Ramdev, Monckton, the Morgellons victims 1, and even Randi.

Stories work against truth. They operate with the machinery of prejudice and

distortion. Their purpose is not fact but propaganda. The scientific method is the tool that humans have developed to break the dominion of the narrative. It has been designed specifically to dissolve anecdote, to strip out emotion and leave only unpolluted data. It is a new kind of language, a modern sorcery, and it has gifted our species incredible powers. We can eradicate plagues, extend our lives by decades, build rockets and fly through space. But we can hardly be surprised if some feel an instinctive hostility towards it, for it is fundamentally inhuman.

Storr is seriously out of his depth on the science: he is always at least second-hand from the evidence (when interviewing researchers), and often third-hand (most of his citations are pop science books), and so several chapters suffer from journalism's classic problem, false balance. The reason this isn't a call to shut the book is because he doesn't spare himself, states this repeatedly - and this is in fact the theme of his book: that almost all of us are unable to infer the truth about a shocking diversity of things. 2 Without testimony, without Google, we are revealed as ignorant and helpless apes.

For instance, the Skeptics he encounters are also out of their depth, and deserve calling-out. No one is past the need for doubt.

I am surprised, for a start, that so few of these disciples of empirical evidence seem to be familiar with the scientific literature on the subject that impassions them so. I am suspicious, too, about the real source of their rage. If they are motivated, as they frequently insist, by altruistic concern over the dangers of supernatural belief, why don't they obsess over jihadist Muslims, homophobic Christians or racist Jewish settlers? Why this focus on stage psychics, ghosthunters and alt-med hippies?

During our conversation, I asked Randi if he has ever, in his life, changed his position on anything due to an examination of the evidence. After a long silence, he said, 'That's a good question. I have had a few surprises along the way that got my attention rather sharply.' 'What were these?' I asked. He thought again, for some time. 'Oh, some magic trick that I decided on the modus operandi'... 'So you've never been wrong about anything significant?' 'In regard to the Skeptical movement and my work...' There was another stretched and chewing pause. He conferred with his partner, to see if he had any ideas. 'No. Nothing occurs to me at the moment.'

That's not how memory works though, is it? Storr is too literal-minded and prosecutorial ("I have been looking for evidence that James Randi is a liar"). When Randi corrects himself in the course of a sentence ("I didn't go to grade school at all, I went to the first few grades of grade school"), Storr leaps on this as a serious contradiction rather than just the patchy nature of speech. He talks about his emotional bias against scepticism - but he still leaves in this idiot journo behaviour, the uncharitable coaxing out of flaws.

These chapters were a good ethnography of 'traditional' (nontechnical) rationality. But Storr doesn't know about the other kind (which foregrounds all the cognitive

biases he is so struck and scarred by), so his conclusion about rationalism is completely awry.<sup>3</sup> Disillusioned with particular Skeptics, he reacts by throwing away scepticism:

For many Skeptics, evidence-based truth has been sacralised. It has caused them to become irrational in their judgements of the motives of those with whom they do not agree... This monoculture we would have, if the hard rationalists had their way, would be a deathly thing. So bring on the psychics, bring on the alien abductees, bring on the two John Lennons – bring on a hundred of them. Christians or no, there will be tribalism. Televangelists or no, there will be scoundrels. It is not religion or fake mystics that create these problems, it is being human. Where there is illegality or racial hatred, call the police. Where there is psychosis, call Professor Richard Bentall. Where there is misinformation, bring learning. But where there is just ordinary madness, we should celebrate. Eccentricity is our gift to one another. It is the riches of our species. To be mistaken is not a sin. Wrongness is a human right.

The American title, “Heretics”, is fitting in a few ways: Storr sees these people as persecuted underdogs, he likes many of them, and so he focusses on the arrogance and bias of the - however correct - mainstream figures dealing with them. They certainly have a holy madness, of crying out despite knowing they will be ostracised.

Over the last few months, John E Mack has become a kind of hero to me. Despite his earlier caution, he ended up believing in amazing things: intergalactic space travel and terrifying encounters in alien craft that travelled seamlessly through nonphysical dimensions. And when his bosses tried to silence him, he hired a lawyer. He fought back against the dean and his dreary minions. He battled hard in the name of craziness...

David Irving is interesting in this regard: he does not act like a fraud (e.g. he sues people for libel, even though this brings intensive scrutiny of his research), but rather a sort of compulsive, masochist contrarian. Stranger still, his (beloved) family were all solid anti-Nazi soldiers in WWII. (Storr contorts himself to explain Irving’s identification with Hitler as due to their sharing an admiration of the British forces (...)) Storr’s awful experience on a Vipassana retreat is a vivid example of the Buddhist dark night of the soul. We don’t know what fraction of people suffer terribly from meditation, but despite its cuddly image, there’s surely large overlap with the 8% of people who are clinically depressive and/or anxious. The chapter on psi does not represent the state of evidence properly - perhaps because one of his proof-readers was Professor Daryl Bloody Bem.<sup>4</sup>

---

The ending is stirring but tilts over into foolish relativism:

The Skeptic tells the story of Randi the hero; the psychic of Randi the devil. We all make these unconscious plot decisions... We are all creatures of illusion. We

are made out of stories. From the heretics to the Skeptics, we are all lost in our own secret worlds.

But the question is to what degree! And the degree of lostness, of inverse rationality, varies by many orders of whatever magnitude you wish to pick. Storr's disquiet at the sheer power of cognitive bias, and the systematic failures of yes/no science (that is: statistical significance rather than effect size estimation) is well and good. Gelman:

I think ‘the probability that a model or a hypothesis is true’ is generally a meaningless statement except as noted in certain narrow albeit important examples.

Storr's humane approach is certainly bound to be more compelling to mystics and flakes than e.g. deGrasse Tyson's smug dismissals. But Storr is scared of grey, of the fact that doubt is only reducible and not eliminable. This is because he doesn't know our beautiful, fallible weapons: probabilism, inference, optimisation, Analysis, computability.

I recommend Elephant in the Brain or Rationality from A to Z instead as an approach to the vital, dreadful side of cognition (including advice on how to avoid being a fake, partial, traditional sceptic); they have less angst and false equivalences, and were written by people who understand the balance of evidence.

Actually that's too strong; I am frustrated with Storr because he is so similar to me, except he doesn't grasp that the technical is the path out of (many) biases. There's a lot wrong with it and you should probably read it, and how often can one say that?

*Cross-posted from Goodreads.*

Storr is right that skeptics can lack compassion. The “Morgellons” people are victims regardless of what their aetiology turns out to be (mental illness, nerve disorders, tropical rat mites, or yes malicious sentient fibres). At minimum, they are victims of bad fortune and rigid, actually unscientific medical practices. The Lesswrong style of rationalist has less of this problem IMO (more emotional literacy; more Californian).

This is an imperfect system, as it relies on many secondary sources. Moreover, I do not declare myself to be free of the biases that afflict any writer, and I'm certainly not immune to making mistakes. If any errors are noted, or if new findings supersede claims made in the text, I would be very grateful to receive notification via [willstorr.com](http://willstorr.com), so future editions can be corrected.

Storr:

I am concerned that I have overstated my argument. In my haste to write my own coherent story, I have barely acknowledged the obvious truth that minds do sometimes change. People find faith and they lose it. Mystics become Skeptics. Politicians cross the floor. I wonder why this happens. Is it when the reality of what is actually happening in our lives overpowers the myth that we make of themselves? Are we simply pursuing ever more glorious hero missions? . . .

Important caveat to the headline of that linked article, from Gelman:

The only thing I don't like about Engber's article is its title, "Daryl Bem Proved ESP Is Real. Which means science is broken." I understand that "Daryl Bem Proved ESP Is Real" is kind of a joke, but to me this is a bit too close to the original reporting on Bem, back in 2011, where people kept saying that Bem's study was high quality, state-of-the-art psychology, etc. Actually, Bem's study was crap. It's every much as bad as the famously bad papers on beauty and sex ratio, ovulation on voting, elderly-related words and slow walking, etc. And "science" is not broken. Crappy science is broken. Good science is fine. If "science" is defined as bad articles published in PPNAS—himmicanes, air rage, ages ending in 9, etc.—then, sure, science is broken. But if science is defined as the real stuff, then, no, it's not broken at all.

```
<h3>Why listen to me on this topic?</h3>
<div>
  <i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect
    <ol>
      <li>immersion in the field and/or good priors for what makes for an extraordinary book
      <li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes to write a book review
      <li>incredible amounts of argument-checking, which doesn't need domain knowledge
    </ol>
  I always try to do (3) but surely often fail.</i> <br><br><br>
```

In this case: I know a bit about psychology and cognitive science, certainly more than S

</div>

# Letter to my baby brother

Gavin

2018-08-24

1. Maths
2. Sex
3. Money
4. Meaning

High-school maths forces you to sit and pretend to be a computer: read an input (a bullshit problem), learn these three rules (e.g. when you see a problem of x type, go through “SOHCAHTOA” until you find the correct ratio to use) and dumbly output your steps and the answer. There’s no creativity, no link to reality, no room for you to be yourself, no point. It makes you feel awful and incompetent because there’s no way to hide from your mistakes. Your teachers are dried-up inside from teaching boring things they have no control over, to a hostile and bored audience, for decades in a row. Why on earth would anyone *voluntarily* do this to themselves? Well, because high-school maths is nothing like real maths, like independent maths.

- Because maths is at the root of most of progress. We have electricity because of physics and computers because of mathematical logic + engineering. Disease is rare because of epidemiology - a branch of stats, which is a branch of maths - and engineering (applied mathematics). “When will I use this?”“When do you see an athlete using pressups during a match?”
- Because it teaches good thought. When I was 18 I tried to take a shortcut to knowledge, to wisdom. I hated high-school maths, so I ran quite far in the other direction: philosophy and literature. It didn’t work. All the questions I want to answer\* Because it’s beautiful. Maths is vast and clean. It’s totally above politics. “Physics is the science of determining which subset of mathematics the universe respects.”
- Because it opens the best jobs. Maybe it seems petty to you to choose things based on how much money they bring you - and you’d be right! There’s so much more to life. Except, it’s not just money. Because there’s always a shortage of technical people (programmers, statisticians, data scientists, economists), they are forced to treat you really well. I’ve seen both sides of this, as an Arts graduate and later as a software graduate. Before, finding a job was always hard, took months. This is stressful, if your parents can’t support you (and you’re 23 years old, so why should they?).

These days I get three invitations to apply to things a week. Because my current work know that I'm in high demand, they have to make it worth my while with raises and perks: my income doubled (£26k to £55k) in 3 years. Still, though, there's more to life than being comfortable. Where's the adventure? Where's the aid to those in need?

- Adventure: I work in machine learning, which is the insider's word for AI. As such, I'm part of an extremely grand project: the process of humanity creating computers smart enough that we don't have to work, so that one day our descendants can live their lives free from alarm clocks, 2 hour commutes, from meetings, from bs reports, from Powerpoint, from office politics, from being ordered around. I couldn't do this work if I hadn't done maths.
- Altruism: Well, last year I gave £10k to charity. If you give to the best, most scientific charities, then this works out to saving about 3 lives, every year. This is far more important than not eating meat, or not owning a car, or buying FairTrade, or .So, for a small number of people (those who can sit still during the appalling training phase) maths consists in or opens up the most important things in life: truth-seeking, prosperity, and lasting aid to one's fellow humans.

I read "Math with Bad Drawings" myself, and even though I've been trying to learn this stuff for 6 years now I learned so much. He's a teacher, and I hope that you get a sense of what the real kind of maths is like. Some kids get to learn from this guy, and I am so envious.

This is an example of how stats can salve or remove emotional pain. We are naturally dramatic creatures, and we suffer for it.

everyone needs to learn at least one technical subject. Physics; computer science; evolutionary biology; or Bayesian probability theory, but something. Someone with no technical subjects under their belt has no referent for what it means to "explain" something. They may think "All is Fire" is an explanation.

# The ‘original affluent society’ was not

Gavin

2020-05-16

{% include affluence/links.md %}

Hunter-gatherers spent their time in more stimulating and varied ways, and were less in danger of starvation and disease... The average farmer worked harder than the average forager, and got a worse diet in return. The Agricultural Revolution was history's biggest fraud.

- Yuval Noah Harari

Hunter-gatherers emerged from the “Man the Hunter” conference in 1966 as the “original affluent society.” The main features of this thesis now seem to be widely accepted by anthropologists, despite the strong reservations expressed by certain specialists in foraging societies concerning the data advanced to support the claim.

The ‘original affluence’ hypothesis is a smelly mix of 4 claims:

1. Ancient hunter-gatherers worked fewer hours than agrarians.
2. Ancient hunter-gatherers worked fewer hours than contemporary industrial people.
3. Ancient hunter-gatherers had greater welfare than agrarians.
4. Ancient hunter-gatherers had greater welfare than contemporary industrial people.

Claims (1) and (3) are plausible. (4) is hard to test (though I will note in passing the incredibly short lives of hunter-gatherers - 54% of people dying before 15, and 80% before 40). Let’s look at (2), because we have some hope of addressing that limited point.

## How long did hunter-gatherers work?

<https://ourworldindata.org/working-hours#all-charts-preview>

“the simplicity of their lives stemmed from a zen philosophy that, because they wanted little, they effectively had all they needed.”

"foraging model that combined deep environmental confidence, a lack of materialism, low population density, egalitarianism, lack of territoriality, minimum storage, and an easy flux in band composition"

There are two lines of evidence about our original society, pre-agrarian hunter-gatherers: archaeology ("paleopathology") and studies of contemporary hunter-gatherers (6000 years later, usually with frequent trade with agrarian and industrial society).

However, Gregory Clark, a great scholar, says they were low-hours.  
<https://www.jstor.org/stable/2566627?seq=1> Farewell to Alms

Lee's numbers explicitly counted only the initial foraging of the mongongo nuts, i.e. none of the food processing, firewood gathering or tool maintenance. After adjusting for these, the average !Kung work week is at least 50 hours and probably more. See:

<http://www.rachellaudan.com/2016/01/was-the-agricultural-revolution-a-terrible-mistake.html>

## How long do we work?

```
<h3>Who's 'we'?</h3>
<div>
    Much of the world is still rising out of Victorian horror.
    Some of the prosperity of the rich world depends on the heavy labour of those in poor co
    <br><br>
    I'm not doing much moral accounting in this piece, but the welfare case does depend on t
</div>
```

<https://www.jstor.org/stable/3631086?seq=1> <https://news.ycombinator.com/item?id=17879759>  
[https://en.wikipedia.org/wiki/Working\\_time#/Hunter-gatherer](https://en.wikipedia.org/wiki/Working_time#/Hunter-gatherer)

VS: "Nowadays, food 'collecting', food processing, DIY (taking care of tools) are not counted into working time but come extra and are taken from 'free' time. And they amount to 1, perhaps 2, working day equivalent per week."

VS VS Sure: "40 hours a week" for moderns is also an underestimate.\* But the figure everyone goes around repeating about hunter-gatherers is a more dramatic underestimate. The accurate !Kung estimate is 48-56 hours spent on these things: so we seem to be about the same, but with massively improved quality, cost, and nutrition for us. The debunking applies to the claim that they had more leisure than us. Lee says: "work week... of 2.4 days per adult... [the bushmen] appeared to enjoy more leisure time than the members of many agricultural and industrial societies."

- Industrialised-world cooking time per week seems to be about 6 hours.  
<https://www.statista.com/statistics/420719/time-spent-cooking-per-week-among-consumers-by-country/> Difficult to find average shopping time,

but call it 5-7 hours a week. <https://www.statista.com/statistics/521924/time-spent-household-shopping-countries/>

How could this be? How could people make sweeping, highly counterintuitive claims off the back of a misreading of two tiny studies? How could the mistake persist for 50 years (and counting)?

{% include affluence/keynes.md %}

## See also

- Kaplan on the darker side
- Rachel Laudan's great piece started me on this trail.
- My worries about cultural anthropology in general.
- On Stress

UK part-time median wage per hour = £9.94 in 2019.  $9.94 * 15 = 149$  a week  
149 a week \* 50 weeks = £7455

1928 average UK working hours: 47 per week Projected ceiling: 15 per week  $1 - 15/47 = 68\%$

1928 average UK working hours: 47 per week 2020: 37 per week  $1 - 37/47 = 21\%$

<h3>Discourse on me</h3>

<div>

I share Keynes' instincts - not as a theory of what other people generally want or should

<!-- -->

I've been tested: I was once unemployed for a long 6 month stretch. Money aside, I didn't

<!-- -->

When I see the point, I work relentlessly. When I don't, it is hard to wake.

</div>

# Hardening the browser

Gavin

2018-09-01

{% include browser/links.md %} {% include js/lazyFrame.html %}

It's now common knowledge that we're being watched online, by a thick mix of nation-states, private companies, and criminals. They sometimes do worse than watch. What do we do? Should we care?

It's not clear what the probability of having your password leaked in a breach / having your email read / having your laptop being remotely wiped (unless you pay the creator Bitcoin) is. But something like this will probably happen to you in your lifetime, so I would take 10 mins to mitigate them now.

There is no absolute security; it's always partial and relative to a goal. This guide is aimed at "*not losing control of your accounts, not being surveilled by companies or criminals, not having your online banking subverted, not getting infected by ransomware or whatever*". It's strictly for people with average risks: not that much money, not much tech cred, not much sensitive information to protect.

"Wait, isn't that your own computer -"

On a lighter note, security is an amazing way to learn about how the internet actually works. It's a lot easier to remember the dozens of abstract systems involved when you can think, smugly, "*And I've plugged that gap with this mitigation, and that one, and that one...*"

Most of this article assumes you're using Firefox, because Chrome is itself an attack. That is, it protects you very well against everyone except Google.<sup>2</sup> It's not a big deal compared to the other parts of this list, you'll just need to find alternatives to the add-ons I recommend.

<h3>Ugh factors and tail risks</h3>

<div>

Why care about this? Besides mere trust in one's hardware, or a mere preference not to b

Only half of humanity are online at the moment; a single script-kiddie troll can do quite

### ***First: password hygiene***

#### **Attack: password cracking**

If people hack a website you're registered on, they could easily get the encrypted 'hash' of your password even if the site owners do everything right. These can eventually be brute-force decoded, and then they have your password. To prevent this common occurrence, we need our passwords to be very long (16 characters +) and have no English words. You also want a different password for each site, so that one brute-force doesn't open up all of your accounts at once. So, easy!: We want passwords that are too hard to remember, and we need to never reuse any of them.

Mitigation: A 'password manager', for instance the free, open-source, cross-platform KeePassX. Keep the database file on several devices, and on a thumb drive, and an offsite. Can put it in the cloud if you think you're likely to lose those. LastPass and 1Password seem fine, maybe a bit slicker and more friendly, but they cost.

You can also sign up to the security researcher Troy Hunt's notification tool: whenever a big leak becomes publicly known, he'll scan it for you and email you if you're in it.

#### **Attack: password phishing**

People can create convincing clones of websites just so you give them your password freely. (This isn't just about human inattention: attackers can register urls which look exactly like the real one).

Mitigation: Password manager / no password reuse.

Real mitigation: Two-factor authentication (2FA) everywhere you can, e.g. via a Universal device like Yubikey. If the site doesn't ask you for the access code from your phone when you sign in, you *immediately* change your password (from the top search result for that site).

(Sadly, SMS confirmation is relatively easy to subvert, so you should use a smartphone. An open-source 2FA app, Authenticator, is coming along though.)

Cognitive burden: once you have the Master passphrase memorised (not hard, give it a couple days): much less than remembering 40 different passwords.

In early 2019, there was splashy media coverage of a vulnerability in all the big password managers. It's true that decoded passwords you've used during a session can persist in your RAM; however, it's of little importance, since if an attacker is in a position to read arbitrary things off your RAM, you are already as screwed as you can be. (KeePass was the least vulnerable manager, incidentally.)

### **Then: Browser**

**Attacks: IP tracking, unencrypted traffic, ISP logs, public wifi spoofing, geo-locking, national bans**

In many jurisdictions (e.g. UK) your internet provider is legally required to record some info about your browsing. In others (US) they do it apparently for kicks. They also implement court orders banning particular sites. Some content is only licenced for computers in particular locations. *And* using public wi-fi (airports, coffee shops) is also extremely insecure without extra encryption.

Partial mitigation for all these: a VPN. This is highly imperfect but not as useless as this guy thinks. They at least have some incentive not to log you: no one will use a VPN which is known to log. I use PrivateInternetAccess; you can check the technical and legal specs of dozens of VPNs here or just get good live recommendations here. \$30 a year. Do not use free ones.

The other problem a VPN solves, and solves optimally, is internet requests sent by non-browser apps on your machine. If you use e.g. Linux's built-in VPN client, everything goes through it.

You should not consider this strong privacy, cover for anything illegal. It's just the minimum required to *do it* in the first place nowadays.

(NB: Modern browsers have a useful thing called WebRTC. It leaks your IP though, so if you really want to hide that you'll need to go into `about:config` and set `media.peerconnection.enabled` to false. uBlock seems to fix this too.)

### **Attack: Man-in-the-Middle**

Even when the URL is real, vulnerabilities in the original internet protocol mean people can sometimes insert themselves inbetween your data and the receiving site. This is lethal (think online shopping, online banking). This add-on prevents this where it can.

(Previously I recommended HTTPS Everywhere, but that depends on a big central database and sends all your requests there, which - though they're lovely people doing this for excellent reasons - is somewhat counter to the spirit of the thing.)

### **Attack: Tracking and fingerprinting**

There are many, many ways to identify someone on the internet, from obvious ones like IP to desperately cunning ones like making your graphics card identify itself or spotting you based on the way you type. Here are some reputable add-ons for Firefox that kill most of this:

- NoScript. Disables all Javascript by default; this stops 90% of attacks and trackers. It is the most important, but also the most costly in time by far. It remembers which sites you let through though, so after about two

weeks this burden becomes negligible. NoScript has a bunch of other cool protections too, vs XSS, clickjacking...

- Privacy Badger. Watches for processes sending information about you. Trying to fix sites' incentives by not blocking sites whose content actually obeys your Do Not Track settings. Seems to cover the use case for both Disconnect and Ghostery.
- DuckDuckGo. The zero-tracking search engine. Not as good as Google, but it includes a built-in “use Google safely” command.
- Cookie Autodelete. Deletes cookies (files placed on your computer to identify you) when the tab is closed. Good compromise. 3
- Facebook Container. Facebook follows you around the internet to a surprising degree - e.g. any time you see a “Login via Facebook” button or a social-media bar with Share buttons, FB polls its cookies to tie you to that site. They sell this to advertisers, which explains the eerie echo effect of your searches. This official Mozilla extension puts the FB cookies in a “container”, an impenetrable box, stopping the passive tracking (they’ll still get you if you click the buttons).

I imagine everyone who will already has, but: consider quitting Facebook or neutering it. You can download all your data from them here, with like a week of waiting.

### Attack: Ads

This one is arguable: the current web economy couldn’t exist without ads. My response is to precommit to using any micropayment solution that people can get to work. Also to actually buy things from creators I like. In the meantime no-one gets to spam me with gigabytes of ugly unwanted content and follow me around.

But besides being ugly, besides following you without your consent, they take your time. Two-thirds of all script execution time is due to third-party scripts, mostly ads and trackers. My own network analytics say that 15% of all my requests are to ad servers. This is hours of your life per year. 1

Everyone knows this solution, but a better solution takes a bit of work:

The best thing to do against ads, at present, is a Pi-hole, a tiny DNS server in your house. This stops ads at the source, for every device in your house at once. You can get a Raspberry Pi for \$30, and it takes about 30 mins to set up as a Pi-hole.

Another benefit of doing this at the router level is that it gives you a nice (rudimentary) network dashboard:

Because the internet is a Red Queen hellscape, we should expect this to gradually stop working over the next few years. Ads can avoid a DNS block in a variety of ways, up to and including them implementing their own custom domain-over-HTTPS protocol. La lotta continua.

### **Attack: email surveillance**

Not a lot you can do, short of undertaking the 100-hour hell of running your own mail server. Try a Swiss company, e.g. Protonmail (they have no public data-sharing agreement with the Five Eyes and constitutional protections for foreigners).

Important caveat: you *really* need to backup your Protonmail password well: If you lose it and reset, you lose your email history. This is the harsh nature of strong security.

Because of the encryption we use to protect your data, resetting your Login password in ProtonMail is different from other, less secure email services. Your password is used to decrypt your emails, and we do not have access to it. Therefore, if you forget your password, you will lose the ability to read your existing emails.

PS: Hotmail and Outlook have been a dumpster fire for many years.

### **Attack: deanonymisation**

No whois entry on your sites. People will try and charge you \$10 for this but it is mandated by GDPR so shop around.

### **Attack: tracking over CDNs**

A new clever attack: identifying you by your repeat requests to a public Content Delivery Network. This add-on DecentralEyes foils this by keeping a copy of commonly-used files in your cache.

Total annual cost: \$45 (\$40 VPN, \$2 usb drive for your password DB + maybe \$4 electricity for the Pi-hole.)

Daily time cost: Net time saving? You'll take a minute a day adding new sites to your NoScript list. And Captchas pop up more often without cookies. But the Pi-hole speeds up your internet by ~10% by not loading ads. And once you get the KeePass keyboard shortcuts in your muscle memory it is faster than typing. So net gain.

---

### **Add-on risk**

Whenever you install a browser add-on, you're allowing unknown code to execute on your machine, behind NoScript. Processes are “sandboxed” in modern browsers - that is, browser malware is unlikely to break into your main OS account - but this is still a risk.

Worst is when someone replaces an honest add-on with a malwared version. This is not hypothetical: for example, part of the Python central package repository was subverted in 2017. And it can take months for someone to notice this.

However, you can be very confident in EFF and Mozilla products - HTTPS Everywhere, Privacy Badger, Containers - and relatively confident in popular open-source add-ons like NoScript, Cookie-Autodelete, uBlock, especially if you built from source.

Still, lean toward avoiding others.

---

#### **More things you could do:**

- Get Linux (99%+ of malware doesn't work on it, and there's strong prevention of state backdoors and 'security through obscurity' zero-days).
- Turn off these Firefox configs.
- "Hacker tape" (putting a removable cover over your webcam) is a successful meme. Good for it! But an even more significant risk is the built-in mic: your unguarded speech is a much more high-res thing to use against you. (Imagine your employer hearing you complain about them to your partner.) One solution is leaving a 3.5mm jack plugged-in, with the wire trimmed off (and the wires taped-up separately to prevent a short circuit!) - but this is still software-mediated rather than hardware, and so could conceivably be bypassed.
- Add an additional keyfile for Keepass, on a USB. This is too far for me. You'd want it attached to your body.
- Tor. Slow!
- CanvasBlocker: people can get a wee bit of identifying info from spying on your GPU and screen specs.
- Airgapping one of your computers.
- ClearURLs (truncate the identifying info from the end of your links).
- CSS Exfil Protection (yet another graphical fingerprinting technique).
- Consider not using Chinese hardware.
- Consider not using American hardware.
- Consider not using Kaspersky (sad - seems to have been involuntary aid to Putin's people).
- Two-factor authenticated bank.
- RandomUserAgent: changes the device and browser you're reporting, at random. Sometimes breaks things.
- Store a PGP key somewhere public (e.g. Keybase): makes it possible to authenticate yourself without identifying documents. (Softening the blow of identity theft, preventing chronic lulz).
- Faraday wallet for phone and contactless card. Obviously this prevents all incoming calls too.
- Life / work separation. Never shop at work, never work on your home computers. This makes two of you, with two different attacks (and sets of attacks) needed.
- *Phone*: The iPhone's encryption has been defended in court against heavy pressure, but also subverted by commercial tools. The Librem 5 will be

better on many axes - hardware control, OS security, supply chain ethics - but is unlikely to do better in crypto.

- Against reward hacking (that is, being distracted with push notifications and infinite feeds): Just don't have a smartphone, or keep it in your bag and use a dumbphone for interpersonal alerts. Also ImpulseBlocker.
- 

Here's a couple of good tools for seeing if this does the trick.

Note that you're not going to stop any nation-states except via perfect paranoia, the kind which makes the above look sloppy and carefree. Luckily, that effort is not worthwhile for almost anyone.

## See also

- Your Computer Isn't Yours
- Violet Blue on resisting tracking, surveillance, devices.
- F-Secure on the whole deal.

{% include browser/foots.html %}

# Bad introspections

Gavin

2019-08-17

{% include intro/links.md %}

Forces of digestion and metabolism are at work within me that are utterly beyond my perception or control. Most of my internal organs may as well not exist for all I know of them directly, and yet I can be reasonably certain that I have them, arranged much as any medical textbook would suggest. The taste of the coffee, my satisfaction at its flavor, the feeling of the warm cup in my hand — while these are immediate facts with which I am acquainted, they reach back into a dark wilderness of facts that I will never come to know... Where am I, that I have such a poor view of things? And what sort of thing am I that both my outside and my inside are so obscure?"

— Sam Harris

A popular method for finding things out is introspection, first-person reflection on your current mental content. Many of the rankest falsehoods were born this way - from absurd religious dogmatism, to psychoanalytic fairytales, to everyday delusions about one's motives and qualities. It has surged in the last decade, under the modest and retroactively scientific branding "mindfulness".

As usual I'm suspicious. Knowledge comes from perception (sometimes), reason (sometimes), memory (sometimes), testimony (sometimes) - the contribution of this other thing is unclear.

An empirical argument against introspection is that we've been introspecting for like 200,000 years (or, properly, for 3,000) and yet we didn't know very much about our minds until about 150 years ago, when we started to use other methods. (Against this, you could separate out two goals for psychological work - *personal instrumental* ones and *general scientific* ones - and then argue that without introspection we'd have been even worse at the first goal, over our species' history.)

## Bad kinds of introspection

### As backdoor to objective reality

- Revelation or kashf. Mistaking a hallucination for contact with ultimate reality.
- Self-evident inference e.g. Descartes has this regrettable habit of leaping from “clear and distinct” (inconceivably-false) ideas to big synthetic claims. He thought he could establish the existence of God by just noticing that he has an idea of god, a perfect thing.
- Inference to one’s past If you use your current feelings as evidence for surprising claims about your distant past. e.g. From introspective things that a patient told someone else, Freud inferred that her serious respiratory/neurological illness was *caused* by her resenting her father for his terminal illness. The history of psychoanalysis (cold-reading) from this Patient Zero on is full of this kind of thing, but the worst single event in it is maybe the lingering false memory craze of the 90s, which harmed thousands of people by leading them to make horrible mistakes about their early childhood, based on Freud’s false ideas about repression. (To what extent is predatory/collaborative delusion even introspection? I don’t know.)
- Inference to deep time Jung’s idea of the collective unconscious is a mashup of a scientific hypothesis (“humans all share the following specific ideas as a result of our common ancestry”) and a completely mad telepathic world-mind thing. Something like this might be possible - just not with this little data, or this method, or this investigator, or this entire worldview.

### **As backdoor to subjective reality**

- Inference to the unconscious mind e.g. People insist on trying to find deep truths about the unconscious mind via dream interpretation, expending lots of ingenuity on what might well be a semi-random byproduct of long-term memory encoding.
- Inference to latent identity It’s now common to identify what you *feel like* with what you are. This has good and bad sides, but in general the idea of a personal essence (as opposed to a personal family-resemblence of contingent properties) is false, and might imply a bad epistemology. (False since you would be a different person if your circumstances changed, even as little as “who you are currently talking to.”)

### **As waste**

- Sitting with your eyes closed telling yourself you’re not thinking. Which is what many ‘meditation’ sessions probably are.

The common failure above is taking introspection too seriously. If you’re doing it for fun or catharsis, and manage to prevent it leaking into your beliefs, then good for you. It’s an art in fact - consider improv, freestyling, automatic writing, internal family systems. I’m only hostile to the epistemic side.

## Phenomenology & mindfulness

Phenomenology is a sort of philosophy that focusses on introspecting ‘structures’, facts about consciousness. (I am frustrated that I can’t find a list of facts they claim to have found, in their century of striving, but not surprised.) This is as opposed to psychophysics, the cool quantitative study of stimuli and their mental results. To me, philosophy is the impersonal attempt to be maximally pedantic, but who knows, maybe it pays to be pedantic about subjective experience.

And mindfulness is sanitised religious contemplation. (Then there’s ‘Focussing’.)

I don’t know very much about either, but some normally critical people I admire think they are *very* good, so they might not be bad introspections.

### <h3>Experimental introspection</h3>

<div>

There may be non-propositional, non-procedural knowledge. It wouldn't be surprising - th

How to test this? If the epistemological side of focusing was real, what would be differ

I don't know. We are too skilled at deluding ourselves. But it would be pretty easy to r

</div>

### <h3>Open questions</h3>

<div>

Why should there be any therapy that works in general?<br><br>

Grant that there is bodily knowledge; where is this knowledge stored? The enteric nervou

Why should introspection work? Theory of mind is for modelling other people so that they

</div>

## Is there good introspection?

Of course; consider what happens when you rate a film you just saw out of 5, or in fact when you give any opinion.

The point is that what you get from introspection isn’t truth, but raw data - data that may need tremendous processing (cross-referencing, explanation in evolutionary or personal-history terms, correction for known biases) to even on average increase your self-knowledge. Also that taking the measurement will alter the mental content, to a possibly useless degree.

Rules of thumb might be: Don’t take it literally; don’t imagine you’re in contact with your unconscious or your essence; don’t generalise, even to your past or future *self*; use it as at most weak Bayesian evidence about the idea.

Justified uses for introspection, for me:

- *Belief propagation.* It seems to help with aligning different parts of the mind, for instance getting my automatic and explicit circuits to pass information. Often a premise will change (“System 2”), without the intuitive associations

changing (“System 1”). (Though I endorse resisting the inverse changes, where your feelings determine a belief.)

- *Hypothesis generation.* If you don’t know what’s wrong it is obviously helpful to get ideas from an entangled source.
  - *Aid to debiasing.* Noticing is moving things from the periphery of your attention into consciousness, where you can evaluate it. For instance, people often don’t “notice” their own current emotional state in this sense, but that’s vital information if you’re trying to be rational - if you’re feeling threatened by a person or a topic, you’re primed to reject arguments around it. A cue to double-check your reasoning, or to revisit once you’ve calmed down.
  - *Emotional processing.* I don’t know how or when thinking about things makes you feel better. But it usually works for me.
  - *Pretext for deep conversations.* I’ve done a few of these kind of workshops, and every single time I meet really interesting people who are there to open up and talk about fun serious things.
- 

## Related

- Schwitzgebel on our broad ineptitude
- Boring (1953). A history of introspectionism
- Danziger (1980). The history of introspection reconsidered

{% include comments.html %}

# What is best?

Gavin

2019-08-23

Whatever the Theosophical cafés and Kantian bistros say, we're deplorably ignorant of the nature of the good.

– Marcel Proust 1

Not to do good, because I don't know what good is, nor even if I do it when I think I do. How do I know what evils I generate if I give a beggar money? How do I know what evils I produce if I teach or instruct?

• Pessoa

Here's a story. You notice that, though poverty is falling, is scarcer than it has ever been, that many people still live lives of abject poverty. This is clearly bad; everyone agrees. You set out to help however you can.

Say then that you notice that one of the first things people do when they rise out of poverty is to increase their meat consumption, and industrialise (that is, torture) their animals.

WAS net-negative

Climate catastrophe

AI slowdown.

(I don't endorse this model, or maybe with <5% credence, it's just an illustrative example.)

It's not just that what you were doing before is less good than you thought; at each step the sign of the value of your action flips entirely. Up is down, altruism is harm.

<https://forum.effectivealtruism.org/posts/NQR5x3rEQrgQHeevm/what-new-ea-project-or-org-would-you-like-to-see-created-in#too3tJZWC74jzhiSj>

1. Cause-impartiality: to select causes based on impartial estimates of impact.
2. Cause-agnosticism: to be uncertain about which cause is highest impact.
3. Cause-divergence: investments in multiple causes.
4. Cause-generality: doing things that can affect any cause

Moral impartiality

If Earth dies, the entire universe dies?

### **Why x-risk and not cause prioritisation?**

Pragmatic Argument: Only two philosophies imply that the end of the world and the prevention of the future are not terrible:

Nihilism (value is not real) Negative utilitarianism ()

### **Top 10 causes**

1. Risks from advanced AI
2. Risks from biological weapons
3. Risks of fundamental moral error
4. Risks from
- 5.
6. Risks from catastrophic climate change?
- 7.
- 8.
- 9.
- 10.
11. Ending death

### **Cause scepticism**

Naturalism -> Anti-real consequentialism -> EV -> X-risk -> AI safety

The goal is maxipok, “no unalignment”.

### **Ideas**

- 
- Free PIGD for every new parent in the world.

*I owe most of the ideas in this to Anders Sandberg, Toby Ord, and James Dama.*

<li class="footnote" id="fn:1">  
Que les cafés théosophiques et les brasseries kantiennes en prennent leur parti, nous ig</li>

# Consent as conclusive evidence

Gavin

2019-08-09

```
{% assign halluc = "https://en.wikipedia.org/wiki/Deathbed_phenomena#Scientific_evaluation"
%} {% assign dmt = "https://www.bbc.co.uk/bbcthree/article/dd52796e-5935-
414e-af0c-de9686d02afa" %} {% assign pearce = "https://www.physicalism.com/#8"
%} {% assign benatar = "https://maverickphilosopher.typepad.com/maverick_philosopher/2017/12/david-
benatar-on-the-quality-of-human-life.html" %} {% assign frankish =
"https://www.keithfrankish.com/illusionism-as-a-theory-of-consciousness/" %}
{% assign crockett = "https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2726330"
%} {% assign val = "https://en.wikipedia.org/wiki/Valence_(psychology)" %}
{% assign harm = "https://en.wikipedia.org/wiki/Harm_principle" %} {% assign
singer = "https://theapanpsychcast.com/singer-preference" %} {% assign hedge =
"https://thingofthings.wordpress.com/2018/11/14/moral-hedges/" %} {% assign
parl = "http://users.ox.ac.uk/~mert2255/talks/parliamentary-model.pdf" %}
{% assign arousal = "https://en.wikipedia.org/wiki/Misattribution_of_arousal"
%} {% assign parf = "https://philpapers.org/archive/PAROAT-4.pdf" %} {% assign
benth = "http://transcribe-bentham.ucl.ac.uk/td/JB/014/049/001" %}
```

A thought-experiment that shows a problem with naive utilitarianism:

Harry and Tom are soldiers on their way back from a mission deep in enemy territory, out of ammunition - when Tom steps in a trap set by the enemy. His leg is badly injured and caught in the trap. Harry cannot free him from this trap without killing him. However, if Harry leaves Tom behind, the enemy troops will find him and torture him to death... Enemy troops are closing in on their position and it is not safe to remain with the trapped comrade any longer... Tom pleads to Harry: "Please, don't kill me. I don't want to die out here in the field" Should Harry stab Tom in the heart to prevent his suffering at the hands of the enemy?<sup>1</sup>

Naively, Harry-as-Act-Hedonic-Utilitarian should kill Tom: he can't save him, so the choice is between {a fast unpleasant death} and {a horribly slow unpleasant death}. So the point is: doesn't hedonic utilitarianism perversely disregard consent?

Two strong reasons it doesn't:

1. we really shouldn't have much confidence that *any* moral theory of ours is

the right one - and this implies giving weight to other theories, particularly when they strongly conflict with ours. This is enough to not kill Tom.

2. but even on the object level, consent is very important to a hedonic utilitarian: it is the only strong, granular evidence we have about valence. This is also enough to not kill Tom.
- 

I'll elaborate on that, but first some other answers:

### **1. Appeal to downstream harm**

A weak response is that you'll make utilitarianism look really bad if it gets out, which we can expect to have larger bad effects. But we can stipulate this away in a thought-experiment.

(Anyway, people are really good at making viewpoints they don't like look bad, even without actual cases.)

### **2. The Millian patch**

John Stuart Mill's version is one of the most liveable forms of utilitarianism because it adds a second principle to pre-empt this kind of thing:

The only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others.

### **3. Rule utilitarianism**

Maybe the rule "don't kill people against their will" is more than good enough to balance out the disutility of one painful death. It pretty plausibly is, but the question is: does the 'mercy killing' weaken that rule in any way?

### **4. Preference utilitarianism**

That consent is ignored by naive act hedonism is a main selling point of 'preference utilitarianism', the kind that maximises satisfied goals rather than raw feels. This is fine, but it doesn't address the original thought-experiment, which is about the hedonic sort.

(Note that the main proponent of this view has switched to the hedonic view because of even more difficult edge cases.)

### **5. Massive uncertainty**

Those 4 responses are all very well if we assume the theory, but I think the real answer to this (and in fact the answer to whole classes of niggling moral edge-cases) is less presumptuous:

### ***Moral uncertainty* (about the right ethical theory)**

We just aren't sure enough of hedonic utilitarianism to act in ways which contravene other moral intuitions as much as this case demands. The details of moral uncertainty are still being worked out, but the general lesson is to hedge (pick things that accord with all good theories), and to trade with people who have different ethics.

plus

### ***Intersubjective uncertainty* (about the value of others' experiences)**

Self-report (and its subset, consent) is currently *overwhelmingly* the strongest evidence we have for the wellbeing of others.

The only hard reason I have to generalise my own situation-valence pairs to others is the deep similarities of our brains. But we know that people with quite similar brains can still have astonishing variance in preferences - witness kink, or addiction, or free soloing. 2

We know too little about suffering to act drastically according to merely intuitive external judgments of experience quality.

Even if it seems obvious that the physical pain of the dying man outweighs the satisfaction he gets from having his wishes respected, I don't have anything like warrant to act. Whether the benefit to him is pride in defying suffering, or in the deeply altered states involved is moot: it is overwhelming evidence even accounting for the chance of Tom lying.

Let's use the word '*overknowing*' for being confident enough to do something *prima facie* bad. My claim is that we don't even overknow the *sign* of another person's valence, without their honest report.

(This might change a bit when welfare biology advances, giving us another source of evidence about the value of a state, but it'll never be countermanded unless we discover some far-out theory that lets us empirically measure the value of a conscious state.)

This is not a post-hoc fix; Bentham himself said it:

Every person is not merely the most proper judge, but the only proper judge of what with reference to himself is pleasure: and so in regard to pain.

<h3>Subjective uncertainty</h3>

<div>

One possible counterargument to the above is that, as well as me being uncertain, the vi

How does this work? You cannot be mistaken about being in pain, for instance. But you ca

[\[David Benatar\]\(#\)](#) argues an extreme version of this, that people a

I don't know what to make of this - it reminds me of <a href="{{frankish}}>illusionism</div>

*Thanks to Saulius Šimčikas, Jan Kulveit and Hugh Panton for conversations on this.*

(modified from 'The Soldier's Dilemma' here.)

A personal example: Derek Parfit says that the minimally pleasant life, the one only just worth experiencing, is eating potatoes while easy-listening music plays. But I am delighted every time I eat potatoes, and have eaten them every day for years on end in the past.

# ‘Strangers Drowning’ (2015) by MacFarquhar

Gavin

2018-06-28

I don’t know whether there are any moral saints. But if there are, I am glad that neither I nor those about whom I care most are among them... The moral virtues, present... to an extreme degree, are apt to crowd out the non-moral virtues, as well as many of the interests and personal characteristics that we generally think contribute to a healthy, well-rounded, richly developed character... there seems to be a limit to how much morality we can stand.

– Susan Wolf

...the moral narcissist’s extreme humility masked a dreadful pride. Ordinary people could accept that they had faults; the moral narcissist could not. To [André] Green this moral straining was sinister, for the moral narcissist would do anything to preserve his purity, even when doing so carried a terrible price... there was “pseudo-altruism”, a defensive cloak for sadomasochism; and there was “psychotic altruism”, bizarre care-taking behaviour based in delusion... the analyst surmised that the masking of their own hostility and greed from themselves might be one of altruism’s functions

– Larissa MacFarquhar

...we cannot and should not become impartial, [Bernard Williams] argued, because doing so would mean abandoning what gives human life meaning. Without selfish partiality—to people you are deeply attached to, your wife and your children, your friends, to work that you love and that is particularly yours, to beauty, to place — we are nothing. We are creatures of intimacy and kinship and loyalty, not blind servants of the world. 1

– Larissa MacFarquhar

Twelve profiles of recent radical altruists, and the backlash they receive from the rest of us. (^) Besides, MacFarquhar has some deep reflections on the good life and human nature to work through. So: There are people who shape their lives around the need of the world – in particular around strangers who are constantly, in some sense, drowning. This category of person does more than just work a caring job and be dead nice to those around them: instead, their entire lives are dominated by the attempt to do the most good.

### <h3>Profiled altruists</h3>

<div>

A fairly fearless nurse who organised the Fast for Life and trained generations of Nicaraguan nurses, continuing for thirty years despite specific threats to her life by Contras.

A pseudonymous animal rights activist who has rescued or won improved conditions for millions of chickens.

Two early effective altruists, Julia and Jeff, who live frugally and donate more than half of their salaries to the most effective NGOs in the world. They plausibly save 100 lives a year, far more than a doctor or firefighter (even before considering replaceability).

A real Christian, who opened her church to the homeless (over the hostility of her flock) and donated a kidney anonymously.

A charismatic, outcaste social worker and jungle statesman, who created a self-sustaining leper ashram, 5000-strong, out of nothing. Also his equally hardcore descendants.

A Buddhist monk who created the largest suicide counselling site in Japan, stressing himself into heart disease.

The omni-parents of Vermont, who adopted 24 of the least cute and easy children on the lists.

A taciturn altruistic kidney donor.

A burned-out idealist.<sup>4</sup>

(I've compiled data on their nature here.<sup>3</sup>)

MacFarquhar appears suspicious about these people, whose lives are taken over by their morals. She calls them "do-gooders" while admitting the term is dismissive.<sup>2</sup> Even the most humble and quiet do-gooder is, she thinks, making an extremely arrogant claim: that the moral intuitions of the whole species - i.e. family favouritism, supererogation, the right to ignore the suffering of strangers - are totally wrong. She leaves no-one unsuspected.

an extreme morality as Singer's or Godwin's can seem not just oppressively demanding but actually evil, because it violates your duty to yourself. To require a person to think of himself as a tool for the general good could be seen as equivalent of kidnapping a person off the street and harvesting his organs to save three or four lives... even to ask this of yourself seems wrong, even perverted. Impartial, universal love seems the antithesis of what we value about deep human attachment.

But these lives are victory laps: the victory of broad reason over narrow animality. MacFarquhar is more nuanced, less willing to dismiss particularism, nepotism and speciesism – which are together known as common sense. (Though I have

only a mild case of the radicals: for instance, I am mostly immune to misery about the state of the world, and I help my loved ones without much guilt. I'm giving 10% now and 50% eventually, but I am such a bookish scruff that the absence of luxuries does not really cramp my life at all.)

One part of Williams' humanist case against radical altruism has dissolved in the last decade: the idea that single-minded ethical focus must erode your connection to your community. Well, the effective altruists are growing in number and maturity; they offer a deep, global community of at least partially serious people to support and be supported by: and all with the stamp of moral consistency.

MacFarquhar doesn't much like utilitarianism, but she is too moved and impressed with her subjects to take the standard, safe, quietist line (which her reviewers have tended to). Throughout, she presents contradictory philosophical propositions, and makes it difficult to know which she believes; she constantly uses indirect speech and deictic discussion, blurring her voice with the debate at hand. This is, I think, an impressive rhetorical strategy – an “esoteric” one. The book is addressed to common sense readers, but also to our uncertainty and faint guilt; it's dedicated to her parents, but explicitly constructed to bring us closer to the altruists:

I took out all the physical descriptions because if you're looking at someone's physical appearance, you're on the outside. Similarly quotations, which seem as though they should be the most intimate form, because they come directly from the person's mouth. Again, in fact, the only way you hear someone speaking is if you're outside them. So if you translate quotation into interior thought, which simply means taking away the quotation marks and saying ‘he thought’ rather than ‘he said’ – that's a more intimate way of encountering someone.\*\*\*

So *Strangers Drowning* covertly brings us closer to radical altruism. Her task is not to establish their ethical premises, nor to win over new obsessives: instead, she simply shows us their sincerity and incredible effects on the world – and, better, shows the lack of evidence and interpretive charity behind their opponents' aspersions. (This goes for the Freudians, the Objectivists, and the anti “codependency” crowd.) It humanises the threatening side of ultimate goodness. She mostly avoids editorialising about the radicals. But one of her clear conclusions is that these people are not deficient, instead having something most people lack:

What do-gooders lack is not happiness but innocence. They lack that happy blindness that allows most people, most of the time, to shut their minds to what is unbearable. Do-gooders have forced themselves to know, and keep on knowing, that everything they do affects other people, and that sometimes (though not always) their joy is purchased with other people's joy. And, remembering that, they open themselves to a sense of unlimited, crushing responsibility...

The need of the world was like death, [Julia] thought — everyone knew about it, but the thought was so annihilating that they had to push it out of consciousness or it would crush them. She understood, and yet did not understand, why other

people didn't give more than they did. How did they allow themselves such permission? How could they not help?

while also noting that, in general

If there is a struggle between morality and life, life will win... Not always, not in every case, but life will win in the end. Sometimes a person will die for a cause; sometimes a person will give up for duty's sake the things that are to him most precious. But most of the time, the urge to live, to give to your family, to seek beauty, to act spontaneously... or to do any number of things other than helping people, is too strong to be overridden... It may be true that not everyone should be a do-gooder. But it is also true that these strange, hopeful, tough, idealistic, demanding, life-threatening, and relentless people, by their extravagant example, help keep those life-sustaining qualities alive.

An amazing book, anyway: charged, critical, structurally ingenious, and filled with humanity – or, with this other, better thing. [Galef Type: Data 2, Values 2]

“Sedia hujan sebelum payung” (c) Zaky Arifin (2015)

```
<h3>Good riddance</h3>
<div>
    The chapter on the blitheness and cruelty of the psychoanalysts enraged me - all the mon
    <blockquote><table><tbody><tr><td style="width:120px;font-family: 'Copperplate Gothic'">
        <br><td><i>Altruists are bossy, because the urge that is usually behind the fulfillment
        <br></tr></tbody></table></blockquote>
        <br>(My, what rigorous science.) So, here's yet <i>another</i> way I am fortunate to live
    </div>

    <h3>Why listen to me on this topic?</h3>
    <div>
        <i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect
        <ol>
            <li>immersion in the field and/or good priors for what makes for an extraordinary story
            <li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes
            <li>incredible amounts of argument-checking, which doesn't need domain knowledge
        </ol>
        I always try to do (3) but surely often fail.</i> <br><br><br>
    </div>
```

In this case: I have a philosophy degree and have read millions of words about demanding

Note for later the absent quotation marks around MacFarquhar's report of the psychoanalysts' and Williams' positions.

“Do-gooder” is still much better than Wolf's term “saint”, because, as MacFarquhar notes, to call someone a saint is to nullify the challenge of their actions: saints are not just ‘people who do really good things’; they are (thought to be) a different sort of being. Any movement (like EA) which seeks to make radical altruism mainstream has to resist this demarcation and get people to see such

a life as, first, good; then, possible for them; and then reasonable - the sort of thing that people would do if they thought about it more.

Philosophy - e.g. Peter Singer, Will MacAskill, Toby Ord, Mark Lee, Geoff Anders, Stephanie Wykstra - looms large here, in this little corner of the species; larger than organised religion. Since all of the philosophers are from Analytic departments, this gives the lie to the generalised standard criticism of academic philosophy (: that they are fatally detached from the concerns of society, dehumanised, etc).

MacFarquhar's account of Stephanie is misleading: she makes it seem like she has opted for ordinary amoral innocence, where the real Stephanie has taken on an incredibly high-impact job, activism for oversight of pharmaceutical clinical trial data.

# Existential risk as common cause

Gavin

2018-10-17

{% include xriskAll/links.md %}

Imagine someone who thought that art was the only thing that made life worth living. 1

What should they do? Binge on galleries? 2 Work to increase the amount of art and artistic experience, by going into finance to fund artists? Or by becoming an activist for government funding for the arts? Maybe. But there's a strong case that they should pay attention to the ways the world might end: after all, you can't enjoy art if we're all dead.

1. Aesthetic experience is good in itself: it's a 'terminal goal'.
2. The extinction of life would destroy all aesthetic experience & prevent future experiences.
3. So reducing existential risk is good, if only to protect the conditions for aesthetic experience.

The same argument applies to a huge range of values.

1. [good] is good in itself: it's a 'terminal goal'.
2. The extinction of life would destroy [good], and prevent future [good].
3. So reducing existential risk is good, if only to protect the conditions for [good]. 3

Casper Oesterheld gives a few examples of what people might plug into those brackets:

Abundance, achievement, adventure, affiliation, altruism, apatheia, art, asceticism, austerity, autarky, authority, autonomy, beauty, benevolence, bodily integrity, challenge, collective property, commemoration, communism, community, compassion, competence, competition, competitiveness, complexity, comradeship, conscientiousness, consciousness, contentment, cooperation, courage, [crab-mentality], creativity, crime, critical thinking, curiosity, democracy, determination, dignity, diligence, discipline, diversity, duties, education, emotion, envy, equality, equanimity, excellence, excitement, experience, fairness, faithfulness, family, fortitude, frankness, free will, freedom,

friendship, frugality, fulfillment, fun, good intentions, greed, happiness, harmony, health, honesty, honor, humility, idealism, idolatry, imagination, improvement, incorruptibility, individuality, industriousness, intelligence, justice, knowledge, law abidance, life, love, loyalty, modesty, monogamy, mutual affection, nature, novelty, obedience, openness, optimism, order, organization, pain, parsimony, peace, peace of mind, pity, play, population size, preference fulfillment, privacy, progress, promises, property, prosperity, punctuality, punishment, purity, racism, rationality, reliability, religion, respect, restraint, rights, sadness, safety, sanctity, security, self-control, self-denial, self-determination, self-expression, self-pity, simplicity, sincerity, social parasitism, society, spirituality, stability, straightforwardness, strength, striving, subordination, suffering, surprise, technology, temperance, thought, tolerance, toughness, truth, tradition, transparency, valor, variety, veracity, wealth, welfare, wisdom.

So, from a huge variety of viewpoints, the end of the world is *bad*, you say?  
What a revelation!

: the above is only very interesting if we get from “it’s good to reduce x-risk” to “it’s the most important thing to do” for all these values. This would be the case if extinction was both 1) relatively likely relatively soon, and 2) we could do something about it.

### **1) What could kill us all, in the coming century?**

Some big ones are: nuclear winter, runaway climate change, runaway AI, and biological weapons. Regarding (1), 80,000 Hours report an educated guess of the total probability:

Many experts who study these issues estimate that the total chance of human extinction in the next century is between 1 and 20%...  
These figures are about one million times higher than what people normally think.

(And if you think that knowledge of the future is radically uncertain, note you should devote more attention to the worst scenarios, not less: ‘high uncertainty’ is not the same as ‘low probability’.)

### **2) What can we do about it?**

Most of the direct work involves technical research, going into the civil service, or improving the way other big institutions make decisions (e.g. philanthropy, science, NGOs). But anyone can fundraise for the direct work and have large expected effects.

In fact, the amount of funding for mitigating existential risks is a terrifyingly small fraction of total government and charity spending (annually, maybe \$10m

for AI safety, \$1bn-5bn for nuclear security, \$1bn for biosecurity): much less than 1%. Full list here.

Say we did all that. How much would it reduce the risk? We don't know, but a 1% relative decrease per \$1bn spent is a not-obviously-useless guess.

Would this level of mitigation override direct promotion of [good]? As long as you place some value on future generation's access to [good], I think the answer's yes.

So it looks like there's a strong apriori case to prioritise x-risk, for anyone who accepts the above estimates of risk and tractability, and accepts that *something* about life has, or will eventually have, overall value.

---

### Who doesn't have to work on reducing x-risk?

- People with incredibly high confidence that extinction will not happen (that is, well above 99% confidence). This is much higher confidence than most people who have looked at the matter.
- People with incredibly high confidence that nothing can be done to affect extinction (that is, well above 99% confidence).
- Avowed egoists.
- People who think that the responsibility to help those you're close to outweighs your responsibility to any number of distant others.
- People with values that don't depend on the world:
  - Nihilists, or other people who think there are no moral properties.
  - People with an 'honouring' kind of ethics - like Kantians, Aristotelians, or some religions.

Philip Pettit makes a helpful distinction: when you act, you can either 'honor' a value (directly instantiating it) or 'promote' it (make more opportunities for it, make it more likely in future). This is a key difference between consequentialism and two of the other big moral theories (deontology and virtue ethics): the latter two only value honouring. This could get them off the logical hook because, unless "preventing extinction" was a duty or virtue itself, or fit easily into another duty or virtue, there's no moral force against it. (You could try to construe reducing x-risk as "care for others" or "generosity".)

I find it hard to empathise with strict honourers - they seem to value principles, or the cleanliness of their own conduct, infinitely more than the lives or well-being of others - but the intuition is pretty common (at least 30%?).

- People that disvalue life:
  - Absolute negative utilitarians or antinatalists: people who think that life is generally negative in itself.
  - People who think that human life has, and will continue to have, net-negative effects. Of course, deep ecologists who side with extinction would be aiming at a horrendously narrow window, between ‘an event which ends all human life’ and ‘one which ends all life’. They’d still have to work against the latter.
  - Ordinary utilitarians might also be committed to this view, if certain unpleasant contingencies happen (e.g. if we increased the number of suffering beings via colonisation or simulation).
- The end of the world is actually not the absolute worst scenario: you might instead have a world with unimaginable amounts of suffering lasting a very long time, a ‘quality risk’ or ‘S-risk’. You might work on those instead. This strikes me as admirable, but it doesn’t have the kind of value-independence that impressed me about the argument at the start of this piece.
- People who don’t think that probability estimates or expected value should be used for moral decisions. (Intuitionists?)
- You might have an eccentric kind of ‘satisficing’ about the good, i.e. a piecewise function where having some amount of the good is vitally important, but any more than that has no moral significance. This seems more implausible than maximisation.

(That list is long, but I think most of the bullet points hold few people.)

---

## Uncertainties

- We really don’t know how tractable these risks are: we haven’t acted, as a species, on unprecedented century-long projects with literally only one chance for success. (But again, this uncertainty doesn’t licence inactivity, because the downside is so large.)
- I place some probability (5% ?) on our future being negative - especially if we spread normal ecosystems to other planets, or if hyper-detailed simulations of people turn out to have moral weight. If the risk increased, these could ‘flip the sign’ on extinction, for me.
- I was going to exempt people with ‘person-affecting views’ from the argument. But actually if the probability of extinction in the next 80 years (one lifetime) is high enough (1% ?) then they probably have reason to act too (though not an overwhelming mandate), despite ignoring future generations.

- Most people are neither technical researchers nor willing to go into government. So, if x-risk organisation ran out of “room for more funding” then most people would be off the hook (back to maximising their terminal goal directly), until they had some.
- We don’t really know how common real deontologists are. (That one study is n=1000 about Sweden, probably an unusually consequentialist place.) As value-honourers, they can maybe duck most of the force of the argument.
- Convergence is often suspicious, when humans are persuading themselves or others.

<h3>Fates other than death</h3>

<div>

The above talks only about extinction risk, and omits the other two kinds of existential

These have their own wrinkles, but the general point remains: most values are ruined by

</div>

<h3>Criticisms</h3>

<div>

1. "This model assumes utilitarian/long-term ethics, but in fact in the population this  
<!-- -->

I definitely don't assume utilitarianism. I also think I only need a very weak kind of I  
"common" is not meant to mean "universal", but "shareable". how would something be 'shar

<br><br>

<!-- -->

<!-- -->

2. "The structure of the argument would be rejected by contemporary liberals, a dominant  
<!-- -->

If this means Rawlsians, then yes they're not so simple. But they have some principles &

<br><br>

<!-- -->

<!-- -->

3. There are many value systems that place an all-else-equal value on survival, in the a  
<!-- -->

Any ideology plus a proper long-term perspective should be much less willing to make tha

</div>

For example, Nietzsche said ‘Without music, life would be a mistake.’

Steady now!

I think I got this argument from Nick Bostrom but I can't find a reference.

# ‘Blindsight’ (2006) by Watts

Gavin

2018-10-26

```
{% assign anos = “https://en.wikipedia.org/wiki/Anosognosia” %} {% assign  
readi = “https://en.wikipedia.org/wiki/Bereitschaftspotential#BP_and_free_will”  
%} {% assign intrusion = “https://thepsychologist.bps.org.uk/volume-25/edition-  
7/neuroscience-soul” %} {% assign theo = “https://en.wikipedia.org/wiki/Neuroscience_of_religion”  
%} {% assign recon = “https://en.wikipedia.org/wiki/Reconstructive_memory”  
%} {% assign chin = “https://en.wikipedia.org/wiki/Chinese_room#Replies” %}  
{% assign psy = “https://en.wikipedia.org/wiki/Psychopathy_in_the_workplace”  
%} {% assign fur = “https://en.wikipedia.org/wiki/Further_facts” %} {%  
assign illu = “https://en.wikipedia.org/wiki/Eliminative_materialism” %}  
{% assign emo = “http://www.cns.nyu.edu/ledoux/pdf/rethinkingEM.pdf”  
%} {% assign dual = “http://reducing-suffering.org/the-many-fallacies-of-  
dualism/#Libertarian_free_will” %} {% assign han = “http://elephantinthebrain.com/outline.html”  
%} {% assign math = “https://en.wikipedia.org/wiki/Richard_Matheson” %}  
{% assign pinto = “https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-  
6613(17)30190-0” %} {% assign libet = “https://www.discovermagazine.com/mind/libet-  
and-free-will-revisited” %}
```

There is a horror in neuroscience. It isn’t in the paper or the data: it depends on subverting your sentimental sense of self, meaning, will, introspection, spirituality; if you don’t have these, it won’t register. It takes unthreatening academic names like “agnosia”, “readiness potential”, “interhemispheric intrusion”, “neurotheology”, “reconstructive memory”, “semantic externalism”. Also threatening names like “executive psychopath”.

The Blindsight ethos - big damn Gothic fatalist Darwinism - is what you get when you take a traditional worldview (dualism, free will, work as what dignifies life, human exceptionalism, further-fact identity) and slam the disenchanting results of a hundred years of science into it. And then add the century to come’s automation and self-modification.

The book put me in a funk for a week - even though I don’t hold any of the positions it sinks. I suppose this is evidence of Watt’s talent. (“Art is a nonrational tool for persuasion: beware.”) Not the least of its achievements is maintaining its murky nihilism in a world where friendly superintelligences exist.

Because of its actual knowledge, this is weird realism, well beyond Lovecraft’s.

They're coming out of the walls: they're coming out of our best science. The vampires (and, to an extent, the Jovian von Neumann spikefest the plot is about) detract from this deeper horror a bit. Doom; unfixable aberration; people who have warped themselves. If you find Black Mirror too disturbing you might want to give this a miss. Watts even tackles "illusionism" - uniquely I think!

Is it strange that the giant lessons of the cognitive revolution are still rare in fiction? Explanations: simply "the Two Cultures" (i.e. novelists are ignorant); or that novelists are shilling for traditional philosophy, maybe because it sells. (Example of a giant lesson: we do not have introspective access to most of what our brains or minds do, on the level of information processing, action, motivation, or even emotion. You might say Freud found this out - but he didn't use reliable methods, made huge obvious errors, and created a closed unfalsifiable loop and so did not really have knowledge.)

In contrast, Watts knows a great deal, uses it well, and takes seriously what he knows: for instance, readiness potentials are given all the emotional weight they deserve. (At least deserved at the time: They've since been taken down a peg.) This novel has 100 scientific papers listed in the back. The only people who cram quite as many ideas into their books as Watts are Stephenson and Egan.

His scorn for the fumbling entendres of psychoanalysis is also extremely endearing:

According to the experts of that time, multiple personalities arose spontaneously from unimaginable cauldrons of abuse — fragmentary personae offered up to suffer rapes and beatings while the child behind took to some unknowable sanctuary in the folds of the brain. It was both survival strategy and ritual self-sacrifice: powerless souls hacking themselves to pieces, offering up quivering chunks of self in the desperate hope that the vengeful gods called Mom or Dad might not be insatiable. None of it had been real, as it turned out. Or at least, none of it had been confirmed. The experts of the day had been little more than witch doctors dancing through improvised rituals: meandering free-form interviews full of leading questions and nonverbal cues, scavenger hunts through regurgitated childhoods. Sometimes a shot of lithium or haloperidol when the beads and rattles didn't work. The technology to map minds was barely off the ground; the technology to edit them was years away. So the therapists and psychiatrists poked at their victims and invented names for things they didn't understand, and argued over the shrines of Freud and Klein and the old Astrologers. Doing their very best to sound like practitioners of Science.

"So we're fishing for what, exactly? Repressed memories?" "No such thing." She grinned in toothy reassurance. "There are only memories we choose to ignore, or kinda think around, if you know what I mean."

People diss the prose but I think it fits the ethos incredibly well:

We fled like frightened children with brave faces. We left a base camp behind: Jack, still miraculously functional in its vestibule; a tunnel into the haunted mansion; forlorn magnetometers left to die in the faint hope they might not.

Crude pyrometers and thermographs, antique radiation-proof devices that measured the world through the flex and stretch of metal tabs and etched their findings on rolls of plastic. Glow-globes and diving bells and guide ropes strung one to another... Inside each of us, infinitesimal lacerations were turning our cells to mush. Plasma membranes sprang countless leaks. Overwhelmed repair enzymes clung desperately to shredded genes and barely delayed the inevitable. Anxious to avoid the rush, patches of my intestinal lining began flaking away before the rest of the body had a chance to die.

Siri, the sociopath pinhead, is a great character. But also often an infuriating Hollywood Rationalist, and several times he gets the last word, which forces me to suspect Watts. Though the bit where his girlfriend is dying and he refuses to say anything because it would be cliched is clearly intentionally infuriating for the reader. So might be this stupid bit of game theory:

"Well, according to game theory, you should never tell anyone when your birthday is." "I don't follow." "It's a lose-lose proposition. There's no winning strategy." "What do you mean, strategy? It's a birthday." Look, I'd said, say you tell everyone when it is and nothing happens. It's kind of a slap in the face. Or suppose they throw you a party, Chelsea had replied. Then you don't know whether they're doing it sincerely, or if your earlier interaction just guilted them into observing an occasion they'd rather have ignored. But if you don't tell anyone, and nobody commemorates the event, there's no reason to feel badly because after all, nobody knew. And if someone does buy you a drink then you know it's sincere because nobody would go to all the trouble of finding out when your birthday is — and then celebrating it — if they didn't honestly like you. ... I could just... plot out the payoff matrix, Tell/Don't Tell along the columns, Celebrated/Not Celebrated along the rows, the unassailible black-and-white logic of cost and benefit in the squares themselves. The math was irrefutable: the one winning strategy was concealment. Only fools revealed their birthdays.

- this only follows if you have ridiculously strong error aversion, where the value of being certain about others' opinion of you overrules the pleasantness of ordinary interaction.

He mentions (but then averts) the single most annoying error when talking about evolution, which is that "maybe it's better for the p-zombie aliens to take over, since they are clearly fitter than us":

"It doesn't bug you?" Sascha was saying. "Thinking that your mind, the very thing that makes you you, is nothing but some kind of parasite?" "Forget about minds," he told her. "Say you've got a device designed to monitor — oh, cosmic rays, say. What happens when you turn its sensor around so it's not pointing at the sky anymore, but at its own guts?" He answered himself before she could: "It does what it's built to. It measures cosmic rays, even though it's not looking at them any more. It parses its own circuitry in terms of cosmic-ray metaphors, because those feel right, because they feel natural, because it can't look at things any other way. But it's the wrong metaphor. So the system misunderstands

everything about itself. Maybe that's not a grand and glorious evolutionary leap after all. Maybe it's just a design flaw."

(But who cares about fitness? A world without qualia is 'Disneyland without children', valueless by definition.)

His Mathesonian attempt to naturalise vampires is kinda clever (they are a subspecies of cannibal savants), and the exemplar vamp Jukka is one of the best characters in the book - but overall their presence is distracting and off-piste; the right-angles epilepsy thing, the revived-by-corporate-greed schtick, more generally Watts holding forth that corporate culture puts massive selection pressure toward psychopathic nonsentience: all these things jolt me out of his otherwise well-built world.

Besides the vamps, there are other over-the-top ughs. His whole theme of technology as inherently dehumanising, in the style of Black Mirror, is just as cherry-picked and annoying as it always is. The idea that consciousness is unadaptive, and so a one-off aberration in a universe of blind replicators - an idea which steamrolls all objections in the novel - is not obviously true. (For instance, see the global neuronal workspace theory, one of the most striking and elegant ideas I've seen in the entire decade, where consciousness is a vital monitor and integrator of our many brain modules.) But it is true either way that our society is currently 'unadaptive', in the sense of not maximising reproduction. (And thank god for that.)

Wrenching but admirable. Great in spite of itself. For the nonangsty, post-dualist, post-further-fact version read Hanson and Simler instead.

[The novel is free! here]

## Errata for a novel

Like so much of low-power science, some results in this have been overturned or minimised since 2006.

- The corpus callosotomy studies which purported to show "two consciousnesses" inhabiting the same brain (like the character Susan) were badly overinterpreted.
- Readiness potentials seem to be actually causal, not diagnostic. So Libet's studies also do not show what they purport to. We still don't have free will (since random circuit noise can tip us when the evidence is weak), but in a different way.

*Cross-posted from Goodreads. See also my review of Will Storr. See also my list of false or weak psychology claims.*

# ‘Homicide’ (1991) by Simon

Gavin

2018-12-12

```
{% assign hacker = “https://news.ycombinator.com/item?id=17521127&fbclid=IwAR1uFwE3NzWZMLp4v6VyJ
%}
```

A character study of twenty vengeful people and the awful indispensable institution they serve and constitute. The detectives are intelligent and hilarious, but have to navigate two extreme and depressing environments: the streets and City Hall, violence and politics.

Simon was embedded with them, and completely effaces himself, makes this novelistic. We get a glorious outsider view, see things even the detectives don’t:

[The detective] glides past the lockup without looking inside, and so doesn’t see the final, unmistakable expression on Robert Frazier’s face. Pure murderous hate.

He gives a complete chapter to most of the detectives, tracking them through a couple of sordid weeks. They are all distinctive, sharp in different ways, but the approach means it stretches on to 700 pages.

## Blood incentives

The most remarkable thing about it is its informal analysis of the *incredibly* poor incentives the bureaucracy gives the detectives: they’re rewarded for arrests, not convictions, and individually penalised for open homicides. I don’t want to think about what this did to their false arrest rate.

A case in which the pathologist’s finding is being pended is not, to the police department, a murder. And if it isn’t a murder, it doesn’t go up on the board. And if it isn’t up on the board, it doesn’t really exist.

No weight was given to the difficulty of the case - whether witnesses remained at the scene, whether physical evidence existed, whether the weapon was found. All this killed inter-squad cooperation, and led to infighting over dumb luck of the draw.

In human terms, the scene at 3002 McElderry Street was a massacre; in the statistical terms of urban homicide work, it was the stuff from which a detective fashions dreams.

(No other crime counted in the stats, despite Homicide also covering accidental deaths and suicides. So this was an incentive to frame things as e.g. suicide if *at all* possible.)

the chance of actually being convicted of a crime after being identified by authorities is about 60 percent. And if you factor in those unsolved homicides, the chance of being caught and convicted for taking a life in Baltimore is just over 40 percent [in 1988].

You might conclude - falsely - that internal stats are worse than nothing - but only stats as bad as these are. A classic of informal institutional economics.

The nationwide murder ‘clearance rate’ (arrest rate) was 70%. Amazing that it was this high, in that comparatively low-surveillance, low-social-trust place.

The [squad’s] clearance rate - murders closed by arrest - is now 36 percent and falling, a... threat to [Lieutenant] Gary D’Addario’s tenure. The board that gave His Eminence reason for concern six weeks ago has continued to fill with open murders, and it is on D’Addario’s side of the wall that the names are writ in red. Of the twenty-five homicides handled by Dee’s three squads, only five are down; whereas Stanton’s shift has cleared ten of sixteen...

There is no point in explaining that three fifths of D’Addario’s homicides happen to be drug-related, just as seven of those solved by Stanton’s shift are domestics or other arguments... It is the unrepentant worship of statistics that forms the true orthodoxy of any modern police department.

More incentive analysis, on police shootings and the shameful closing of ranks:

In the United States, only a cop has the right to kill as an act of personal deliberation and action. To that end, Scotty McCown and three thousand other men and women were sent out on the streets of Baltimore with .38-caliber Smith&Wessons, for which they received several weeks of academy firearms training augmented by one trip to the police firing range every year. Coupled with an individual officer’s judgement, that is deemed expertise enough to make the right decision every time.

It is a lie. It is a lie the police department tolerates because to do otherwise would shatter the myth of infallibility on which rests its authority for lethal force. And it is a lie that the public demands, because to do otherwise would expose a terrifying ambiguity. The false certainty, the myth of perfection, on which our culture feeds...

There's so much careful and sympathetic detail about the job (and no deep portrait of any suspects), that Simon risks partisanship - writing "copaganda", as internet radicals call it. Anyone who's seen *The Wire* knows this isn't a problem. (He has solidarity with the rank and file, and contempt for the suits.)

for the black, inner-city neighborhoods of Baltimore, the city's finest were for generations merely another plague to endure: poverty, ignorance, despair, police.

Speaking of which: This is not at all made redundant by *The Wire* - the show has an entire pathos-pathetic angle (the anti-authority cop) missing here, and this is more focussed on the law side.

Their humour is fantastically sick. > the application of criteria such as comfort and amusement to the autopsy room is ample proof of a homicide man's peculiar and sustaining psychology. But for the detectives, the most appalling visions have always demanded the greatest detachment...

Someone on Hacker News was up on their high horse about the black humour of medics recently. This strikes me as perfectly backwards. I would much prefer a doctor (or a detective) with a nasty sense of humour: it suggests emotional detachment, so they're more likely to think clearly; and it certainly has a cathartic and bonding role, improving their health and teamwork. This idiotically literal, first-order model of psychology (as if people were so easy to program!) is everywhere, for instance all discourse about fake news, porn, and violent computer games.

The section about the idiocy and arbitrariness of juries is sickening and I recommend that you don't read it if you want to continue thinking well of your society.

The operant logic of a Baltimore city jury is as fantastical a process as any other of our universe's mysteries. This one is innocent because he seemed so polite and well spoken on the stand, that one because there were no fingerprints on the weapon to corroborate the testimony of four witnesses. And this one over here is telling the truth when he says he was beaten into a confession; we know that, of course, because why else would anyone willingly confess to a crime if he wasn't beaten?

The other eight jurors offered little opinion except to say they would vote for whatever was agreed upon... It was the Memorial Day weekend. They wanted to go home... "What brought you all around to first-degree?" he asks. "I wasn't going to budge and that other woman, the one in the back row, she wasn't going to change her mind either. She was for first-degree from the very beginning, too. After a while, everyone wanted to go home, I guess."

The book has aged badly in one way: Simon completely falls for two entrenched bits of pseudoscience: polygraph and profiling . But many people still believe in

these things, and anyway it's a rare lapse of scepticism, for him.

I think this is the first 'true crime' book I've read. I don't know if this is the pinnacle of the genre, or if the genre's better than literary people think.

# Anaesthetatron

Gavin

2014-01-15

```
{% assign weak = "http://en.wikipedia.org/wiki/Philosophical_zombie" %} {%  
assign quo = "https://en.wikipedia.org/wiki>Status_quo_bias" %} {% assign  
machine = "https://en.wikipedia.org/wiki>Status_quo_bias" %} {% assign  
global = "https://en.wikipedia.org/wiki/Global_workspace_theory" %}
```

Suppose there were an experience machine that would give you any experience you desired. Super-duper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life experiences? ... What **else** can matter to us, other than how our lives feel from the inside?

— Robert Nozick

When we talk about the great workers of the world we really mean the great players of the world. The fellows who groan and sweat under the weary load of toil they bear never can hope to do anything great. How can they when their souls are in a ferment of revolt against the employment of their hands and brains?

— Mark Twain

Suppose your workplace installed a machine by the entrance. Say this machine *turned off* your consciousness, leaving the body motive and intelligent, in a weak-AI way. 1 Say your work did not suffer in the least. Say that at 5:30pm, your body steps into the machine again and you are returned to yourself, a little tired but unbored.

This machine is less hedonistic than Nozick's Experience Machine, but still unusual enough to give some people the creeps. How many of us would use the machine regardless, on how many of our days? What does it say about our jobs or our minds that we would?

Sure, this might not be possible if consciousness correlates with some effective neural circuit. But run with it please.

# Perelman

Gavin

2019-01-08

*A review of ‘Perfect Rigour’ (2009) by Masha Gessen.*

```
{% assign kol = "https://www.scottaaronson.com/blog/?p=3376" %} {% assign
drov = "https://en.wikipedia.org/wiki/Aleksandr_Danilovich_Aleksandrov" %}
{% assign emp = "https://en.wikipedia.org/wiki/Soviet_Empire" %} {% assign
what = "https://en.wikipedia.org/wiki/Grigori_Perelman#Possible_withdrawal_from_mathematics"
%} {% assign dfw = "https://www.esquire.com/sports/a5151/the-string-theory-
david-foster-wallace/" %} {% assign mash = "https://www.newyorker.com/contributors/masha-
gessen" %} {% assign yau = "https://en.wikipedia.org/wiki/Shing-
Tung_Yau#Poincar%C3%A9_conjecture_controversy" %} {% assign erd =
"https://www.goodreads.com/review/show/2825487882?book_show_action=false&from_review_page=1"
%}
```

‘Perfect Rigour’ is not so much a biography: instead it’s a study of anti-Semitism in Russia, the viciousness of Soviet academia, and the wonderful subculture that lived uneasily within it. This subculture was the superhuman apolitical dreamland, mathematics. It could only exist because of sacrifices by famous and decent men, Kolmogorov and Aleksandrov. Their selective maths schools seem to have been the only nice places in the entire empire, at least for those with a taste for actual discussion, or unalloyed truth. (It can’t be a biography because the subject refused to talk to her, does things that are very hard to explain, and doesn’t go out much.) Even so, Gessen is well-placed to write this - she was a maths nerd in Soviet Russia around the same time. As far as I can tell (which isn’t very far) her grasp of the maths (one chapter for the crown jewel) is fit for purpose. But Gessen is out to bust Perelman’s reputation for hyper-individualism; so she focusses on the devoted teachers and functionaries that pulled strings to get an abrasive Jew into the heart of Soviet academia, and his incredible luck in starting graduate study just as Glasnost happened. She wants to highlight the poverty of his character - his antisocial withdrawal, his complete and intentional ignorance of politics, his naivete, his savantism. It doesn’t work. Yes, he’s rigid; maybe he is composed of a curiosity, a competitiveness, an ethics, and nothing else (no vanity, humour, romance, charisma, empathy, theory of mind, tolerance, compromise, doubt).

So what? Why does everyone need to be rounded? Does she sneer at athletes, the other people with lives this seemingly contorted and simple? Relatedly,

David Foster Wallace managed to get over himself:

The restrictions on [this pro-tennis player's] life have been, in my opinion, grotesque; and in certain ways Joyce himself is a grotesque. But the radical compression of his attention and sense of himself have allowed him to become a transcendent practitioner of an art – something few of us get to be. They've allowed him to visit and test parts of his psychic reserves most of us do not even know for sure we have (courage, playing with violent nausea, not choking, et cetera). Joyce is, in other words, a complete man, though in a grotesquely limited way.

Gessen is, to be frank, quite cruel: she never passes up an opportunity to mention appearances - that that athletic boy of 1970 is "now an overweight and balding computer scientist", that the house of a man caring for his wife with late-stage dementia is "a messy place, lived in awkwardly" and he himself "similarly unkempt"; that Perelman didn't change his underwear or clip his nails as a teen. This is the shallow side of the New Yorker style on show - or else the malign side of Russian honesty. Either way: no. (Though Perelman would probably approve.)

[Perelman] sounded his voice only if a solution required his intervention; looked forward to Sundays, sighing happily and saying that he could "finally solve some problems in peace"; and, if asked, patiently explained any math issue to any of his classmates though apparently utterly unable to conceive of anyone not comprehending such a simple thing. His classmates repaid him with kindness: they recalled his civility and his mathematics, and none ever mentioned to me that he walked around with his shoelaces undone...

The great mystery, which Gessen understandably can't touch, is why after 36 years of focus he suddenly stopped doing the only thing he'd ever done. How could he? How can that much momentum be shed? What does such a man do next? If you don't care about maths or if you can't abide people being mean to nerds (as both the old apparatchiks and Gessen were) then skip this book.

## Coffins

Valery Ryzhik's story about the evil entrance exam he sat is so, so sad:

"Coffins" were questions specially designed for the Jewish applicants... rejection was administered in a peculiarly sadistic way... if [Jews] succeeded in answering correctly the two or three questions on the ticket, then, alone in the room with the examiners, they would be casually issued an extra question... a problem not merely complex but unsolvable. The examiners would then nail the cover of the coffin shut: the Jewish applicant had failed the exam... "They did not even manage to find a problem I couldn't solve; I sat for three hours after the exam was over, I solved them all, and still they failed me. I was just a boy. I went home and cried."

## Saint Erdos and Saint Perelman

One of the oddest things about Perelman is that he'd disagree with me when I called maths "apolitical". This, combined with his being an actual deontologist in a world of opportunists, maybe explains him turning down a million quid and the highest honours his world can bestow: maths, the least animal and least irrational thing we have, is still too political for him.

[Hamilton] was smiling, and he was quite patient. He actually told me a couple of things that he published a few years later. He did not hesitate to tell me. Hamilton's openness and generosity — it really attracted me. I can't say that most mathematicians act like that.

I personally decided for myself that it was right for me to stay away from verification [of his proof] and not to participate in all these meetings. It is important for me that I don't influence this process."

Perelman told Interfax he considered his contribution to solving the Poincare conjecture no greater than that of Columbia University mathematician Richard Hamilton. "To put it short, the main reason is my disagreement with the organized mathematical community. I don't like their decisions, I consider them unjust."

He mentioned a dispute that he had had years earlier with a collaborator over how to credit the author of a particular proof, and said that he was dismayed by the discipline's lax ethics. "It is not people who break ethical standards who are regarded as aliens," he said. "It is people like me who are isolated... there are many mathematicians who are more or less honest. But almost all of them are conformists. They are more or less honest, but they tolerate those who are not honest."

We asked Perelman whether, by refusing the Fields and withdrawing from his profession, he was eliminating any possibility of influencing the discipline. "I am not a politician!" he replied, angrily.

There was a bit of dishonesty and jostling at the time of the announcement - but nothing compared to any other science, let alone any government. Maybe the protective bubble everyone set up for him was bad for him, because it robbed him of perspective and so made the mild case of fuckery he suffered seem like a complete invalidation of mathematical culture.

But maybe a rigorous rule-based mind would always explode eventually, even if given a scale to measure instances of bias.

The clearest precedent is Paul Erdos: also rude, also monomanaical, also recognisably a saint in some sense.

<h3>Why listen to me on this topic?</h3>  
<div>

```
<i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect  
<ol>  
    <li>immersion in the field and/or good priors for what makes for an extraordinary book  
    <li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes to write the review  
    <li>incredible amounts of argument-checking, which doesn't need domain knowledge  
</ol>  
I always try to do (3) but surely often fail.</i> <br><br><br>
```

In this case: I have a maths degree but basically no comprehension of topology.  
</div>

*Cross-posted from Goodreads.*

# Stephen Jay Gould & the frailty of science

Gavin

2019-01-05

```
{% assign bront = "https://en.wikipedia.org/wiki/Bully_for_Brontosaurus" %}  
{% assign ms = "https://www.nybooks.com/articles/1995/11/30/genes-memes-  
minds/" %} {% assign toob = "https://www.sscnet.ucla.edu/comm/steen/cogweb/Debate/CEP_Gould.html"  
%}
```

Start by listing Gould's virtues: he is passionate about paleontology and paleontologists; contagiously curious about obscure corners of nature and human history; scrupulously fair to the religious and the pre-modern; he is animated by a sense of justice. For an academic, his prose is highly flavoursome and fun. He has a considered opinion about Darwin's handwriting and the meaning of baseball. One of his essay collections was very important to me as a teen, showing me that I could unify truth-seeking and justice-seeking, and with style.

But this is all countermanded: he is just not trustworthy on human topics, and neither on core evolutionary theory, I'm told. From his enormously influential, fallacious dismissal of intelligence research in general and Morton in particular, to his dishonest coup of public discourse over punctuated equilibrium (pushing the flashy and revolutionary version in literary magazines, retreating to minimal and uncontentious forms in the science journals who could actually evaluate it), he muddied the waters even as he brandished real literary talent and noble political intentions. This is unforgiveable: empirical clarity is too rare and precious to sacrifice.

John Maynard Smith:

Gould occupies a rather curious position, particularly on his side of the Atlantic. Because of the excellence of his essays, he has come to be seen by non-biologists as the preeminent evolutionary theorist. In contrast, the evolutionary biologists with whom I have discussed his work tend to see him as a man whose ideas are so confused as to be hardly worth bothering with, but as one who should not be publicly criticized because he is at least on our side against the creationists. All this would not matter, were it not that he is giving non-biologists a largely false picture of the state of evolutionary theory.

Krugman

Gould is the John Kenneth Galbraith of his subject. That is, he is a wonderful

writer who is beloved by literary intellectuals and lionized by the media because he does not use algebra or difficult jargon. Unfortunately, it appears that he avoids these sins not because he has transcended his colleagues but because he does not seem to understand what they have to say; and his own descriptions of what the field is about - not just the answers, but even the questions - are consistently misleading. His impressive literary and historical erudition makes his work seem profound to most readers, but informed readers eventually conclude that there's no there there.

Tooby and Cosmides:

Now, given the foregoing, one is left with the puzzle of why Gould so customarily reverses the truth in his writing. We suggest that the best way to grasp the nature of Gould's writings is to recognize them as one of the most formidable bodies of fiction to be produced in recent American letters. Gould brilliantly works a number of literary devices to construct a fictional "Gould" as the protagonist of his essays and to construct a world of "evolutionary biology" every bit as imaginary and plausible as Faulkner's Yoknapatawpha County. Most of the elements of Gould's writing make no sense if they are interpreted as an honest attempt to communicate about science (e.g., why would he characterize so many researchers as saying the opposite of what they actually do) but come sharply into focus when understood as necessary components of a world constructed for the fictional "Gould" to have heroic fantasy adventures in... "Gould" the protagonist is a much loved character who reveals himself to be learned, subtle, open-minded, tolerant, funny, gracious to his opponents, a tireless adversary of cultural prejudice, able to swim upstream against popular opinion with unflinching moral courage, able to pierce the surface appearances that capture others, and indeed to be not only the most brilliant innovator in biology since Darwin, but more importantly to be the voice of humane reason against the forces of ignorance, passion, incuriosity, and injustice. The author Gould, not least because he labors to beguile his audience into confusing his fictional targets with actual people and fields, is sadly none of these things. Yet in the final analysis, there are genuine grounds for hope in the immense and enduring popularity of Gould. Gould is popular, we think, because readers see in "Gould" the embodiment of humane reason, the best aspirations of the scientific impulse. It is this "Gould" that we will continue to honor, and, who, indeed, would fight to bring the illumination that modern evolutionary science can offer into wider use.

Here is a fictional leaf from Gould's *ad hominem* book, to give you a sense of what he does, at his worst:

Gould is famed for the theory of punctuated equilibrium, which holds that adaption and speciation is not generally a slow, gradual process measurable in tens of thousands of year periods, but instead a rapid response to environmental shocks, measurable in hundred-year periods. The political bias of this theory is too blatant to ignore: as a Marxist, Gould requires that sustainable change be possible by revolution rather than by long accumulation (....)

(For full effect I should now chide him for his genic panadaptationism.) Along with Lewontin and Rose, Gould mediated a huge contradiction in our culture: they allowed the C20th left to feel we were scientific, in our comfortable blank-slatism. That we had already incorporated the deep challenge of evolutionary biology - since these eminent men told us it had no human implications. Read Gould for fun and uplift, but take great care, for he cares about other things more than truth.

(Read Midgley and Singer first if the politics scare you; they might stop you fleeing into Gould's dodgy arms.)

# Mistakes in data science homework

Gavin

2019-04-08

(Takehome exercises)

Forcing everything into classification, particularly binary classification.  
No robustness checks - e.g. of nondeterm algos which narrowly beat  
Non-random train/test split  
Code in a pdf / images  
No code reuse  
More than twice the error of random guessing  
Not normalising for linear models  
Forking paths - multiple comparisons  
Doing PCA but not reducing the damn dimensions  
Interpreting regression leaf nodes as classes  
No library import statements

Things I have seen marking machine learning take-homes:

- No test split.
- Test split but forgot to use it.
- An exquisite description of the mathematical and practical motivation behind PCA, followed by them taking as many components as there were original columns.
- Tree depths of 50 on data with 1000 rows (this is saturated, with one row per perverse leaf, after 10 splits).
- Absolutely everyone uses classifiers on ordinal/interval responses. I seriously want to know why.
- Excellent performance in the Notebook, which disappears once I rerun it (with the same seed). This was probably cheeky .ipynb file hacking.
- Most people forget to normalise the data for linear reg, making everything else in the analysis nonsense.
- No code reuse ever.
- Blithe recommendation of multiple comparisons (cheating).

- Editing the response data, instead of binning or rescaling.

# Important analogies

Gavin

2019-04-08

**Kangaroo jumping on the scale you're weighing a feather on**

**Picking up pennies in front of a steamroller**

Risk insensitive optimisation is not optimisation.

**Chinese robber / Cardiologists**

base rate neglect can stain literally any large group

Chinese robber - Correspondence bias

Cutting the pie, growing the pie

# ‘Ficciones’ (1944) by Borges

Gavin

2019-03-28

```
{% assign mein = "https://en.wikipedia.org/wiki/Meinong%27s_jungle" %} {%  
assign parei = "https://en.wikipedia.org/wiki/Pareidolia" %} {% assign tlon  
= "http://art.yale.edu/file_columns/0000/0066/borges.pdf" %} {% assign bab  
= "http://libraryofbabel.info/Borges/libraryofbabel.pdf" %} {% assign ext =  
"https://en.wikipedia.org/wiki/Semantic_externalism" %}
```

These stories are deeply uncanny, without worshipping mystery. “Tlön” is scarier to me than any of Lovecraft. “Babel” is also horrifying in a way. Borges’ characters are reasoning about the limits of reason. (There is the unearthly drama of higher mathematics in a couple of these.) It manages to be cryptic without being annoying, to use literary gossip and the droning of archivists for art. Some of this is 80 years old, and it’s still completely fresh. He makes literature larger, by bringing in new things - bibliographic minutiae, English department arcana, salon gossip. He writes perfect reviews of fake books. Gushing praise of nonexistent authors draws back the veil (as if our world’s reviewers would say the same things whether or not the authors existed). Borges was not a postmodernist but these have the best of what I take postmodernism to mean: nonliteral play, generative scepticism about sense and reference, language-games. I am often not sure of the significance of Borges’ sentences. But for once the critic’s working assumption of hidden meaning seems sound: if I thought about it, I could find out. And not just in the ordinary way, by projection. I expect to find Borges in them if I try.

## “The Library of Babel”

A banal idea: “language is composite”. Characters go into words into sentences into works into worldviews. Here Borges stretches this fact until you see horror in it, the shock of exponentiation on the tiny scale of a human life.

In the simple idea of mechanically generating all strings of length n=1,312,000, Borges finds a Gothic, claustrophobic closed nightmare. The story is not 8 pages long and contains more thinking than many books.

There exists one truth; there are uncountably many falsehoods; worse, there’s a far larger infinity of nonsense, of things which make sense in no language,

which don't make enough sense to be false, which never will. This is the horror of Platonism or Many-world physics or Meinong: that we could be invisibly boxed-in by garbled infinities, endless keyboard mashing. The “noosphere” - all good ideas plus all bad ideas ever had - is a tiny pocket of meaning in a sea of meaninglessness.

The stunning effect of “Babel” depends on its not being magic, not hand-wavy (merely monstrous, physically impossible for interesting reasons which violate no particular law). Ted Chiang is grasping at a similar titanic scale when he uses an actually alien language to explain variational physics.

Borges was a librarian. But, while he said photogenic things about libraries, he didn't necessarily like being in them. “The Library of Babel” twists that quotation, by imagining an otherworldly library which breaks men just by existing. Sturrock, his biographer:

Borges had some reason to dislike libraries because for nine years “of solid unhappiness”, from 1937 to 1946, he was obliged to work in one, as a quite junior librarian, in order to make money. The cataloguing work he did was futile...

The alphabet used for the Babel books has 22 letters and no uppercase. We could try and look up human languages with that many letters, but better to take this as a hint that our narrator is not us - he can be a total alien, far from Earth, and the exact same library will still confound him the exact same way. The same geometry constrains all minds. Even what seems meaningful need not be, if your sample is large enough:

This useless and wordy epistle [‘The Library of Babel’] itself already exists in one of the thirty volumes of the five shelves in one of the uncountable hexagons - and so does its refutation. (And n possible languages make use of the same vocabulary; in some of them the symbol ‘library’ admits of the correct definition ‘ubiquitous and everlasting system of hexagonal galleries’, but ‘library’ is ‘bread’ or ‘pyramid’ of anything else... You who read me, are you sure you understand my language?)

The narrator says that the fall from his floor of the Library “is infinite” (or indefinite), that the rooms are “uncountable”, but we can do better than this quite easily, from the text. There are  $410 \times 40 \times 80 = 1,312,000$  characters per book. The number of distinct books is thus  $(22 + 3)^{\{1312000\}}$  or ~2 followed by 1.8 million zeroes. (The extra three are space, period, and comma.) It is hard to give a reference for how large this is: if every atom in the universe contained as many atoms as are in the universe ( $10^{80}$ ), and each of the *nested* atoms was a Babel book, this would still contain only a laughably tiny fraction of Babel, less than one googolplexth. There's  $4 \times 5 \times 32 = 640$  books per hexagon, so we need about  $3 \times 10^{1834094}$  room-sized hexagons. This is the full implication of the simple thought “every book of length 1312000”. (Borges notes his own infinity/finity contradiction on the last page, explaining that the Library is unbounded and periodic, a hypersphere.)

It couldn't possibly be even fractionally built. And yet, through maths, it has been built! - "only" implicitly, skeletally. Still counts.

And so a beautiful lesson: think what the incredible feat of writing any book - no matter how bad - actually entails. Our nervous system shields us from Babel, from the larger part of possible meanings and the overwhelming majority of string space. This is an astonishing act, in information terms: the ultimate search, which we succeed at effortlessly, many times a day. Epic achievements in life-giving ignoring.

**The Approach to Al-Mu'tasim**

**Pierre Menard, Author of the Quixote**

**The Circular Ruins**

**The Lottery in Babylon**

**An Examination of the Work of Herbert Quain**

**The Library of Babel**

**The Garden of Forking Paths**

**Funes the Memorious**

**The Form of the Sword**

**Theme of the Traitor and the Hero**

**Death and the Compass**

**The Secret Miracle**

**Three Versions of Judas**

**The End**

**The Sect of the Phoenix**

**The South**

*Cross-posted from Goodreads.*

# ‘Waking Up’ (2014) by Harris

Gavin

2019-04-26

{% include harris/links.md %}

Most people who believe they are meditating are just thinking with their eyes closed.

Forces of digestion and metabolism are at work within me that are utterly beyond my perception or control. Most of my internal organs may as well not exist for all I know of them directly, and yet I can be reasonably certain that I have them, arranged much as any medical textbook would suggest. The taste of the coffee, my satisfaction at its flavor, the feeling of the warm cup in my hand—while these are immediate facts with which I am acquainted, they reach back into a dark wilderness of facts that I will never come to know. . . . Where am I, that I have such a poor view of things? And what sort of thing am I that both my outside and my inside are so obscure? . . . Am I inside my skull? Let’s say yes for the moment, because we are quickly running out of places to look for me. Where inside my skull might I be? And if I’m up there in my head, how is the rest of me me?

A surprisingly humble and sincere book. Some readers feel tricked - that Harris is smuggling in science under soft, false pretences. This isn’t fair; he has done this stuff for decades, visited lamas in Tibet, put in the work. He wouldn’t do so much insincerely; whatever his other failings, he’s actually trying to bridge the two kinds of seekers.

(That said, the cover is a masterpiece of camouflage. Look at the soft colours, the sunny logo, the sans-serif purity, the unthreatening subtitle. Compare his other books!)

Consider all the things people mean by “spirituality”:

1. subjective knowledge of ultimate / immaterial reality
- 1b. gaining supernatural abilities as a result
2. one’s deep moral or existential values
3. personal growth
4. feeling of awe-inspiring beauty
5. introspection; close contact with one’s own “inner dimension”

6. “the ability to step a little back from your emotions and thoughts, observe them as they are without getting swept up in them, and then evaluating them critically”
7. sense of love towards (all) others
8. the quest to see the ego and the self as illusory

With so much popular support - with so much baggage - it's not possible to throw out the word or concept; instead we have to try and reform it. This is Harris' mission - though in fact he focusses almost exclusively on (5) -> (8), the standard Buddhist therapy of not being hurt by distraction, bad luck, frustrated desires, a pesky inner homunculus.

And obviously he rejects (1): we are psychologising the whole thing. Paraphrased: ‘Instead of making you experience reality, meditation lets you experience your mind; instead of strengthening your insubstantial soul, you’re strengthening your mind.’

This is a healthy reconstruction in my view, but it certainly leads him to make controversial claims like “The deepest goal of spirituality is freedom from the illusion of the self”. Metaphysically profligate readers will have no fun here. (But they knew that already.)

How can a scientist (or at least a pro-science talking head) boost a practice whichs aim to reject thought? Well, in most practitioners the rejection is a temporary one. And the trick is to distinguish thinking / experiencing (which are the locus of all value, and of decisions, and of creativity) from identifying with the stream of your thoughts, from being carried away, from being *permanently* distracted.

I’m an unpromising practitioner. For instance, this is kind of my jam. It’s not the indescribability that bothers me - after all, any knowledge-how is indescribable (or rather describable only with millions of parameters). You can accept Hume or Parfit’s reasoning - you can have the propositional knowledge, can know that “there is no self beyond my bundle of experiences”. Meditation is supposed to be the know-how of nonessentialism, the skill of actually paying attention to the implications of this System-2 judgment.

But being ‘nonconceptual’ means no language, no premises, no reason, no jokes, no connection, no comparison. It means using none of my strengths, leaving none of my spoor. On the face of it this is a great loss to me.

I don’t know that I do suffer as a result of identifying with my thoughts; I don’t think that dissatisfaction lurks in every sensation I ever experience or also my whole life in retrospect. But the old claim, similar to Marxist or feminist ‘false consciousness’, is that I am too owned to realise I’m being owned:

beginning meditators... report after days or weeks of intensive practice that their attention is carried away by thought every few seconds. This is actually progress. It takes a certain degree of concentration to even notice how distracted you are.

Freedom from desire sounds much like death to me, for all that Harris and others argue that it can somehow coexist with passion against the suffering of others, with striving to be a better person, with chipping in to the Great Project of discovery, compassion, optimisation. Luckily the two strands of the Buddhist project seem to be separable:

1. really feeling that you are not your thoughts, not a homunculus behind your eyes having them;
2. not wanting things because wanting leads to disappointment.

A consolation: there's a sense in which meditation, introspection and phenomenology are highly, maximally empirical - they involve very close attention and analysis of the raw data. It just happens that the raw data (the sense-data) are irreducible, private, closed, and so not directly a matter for science. Empiricism before science, consciousness without self. I like this part.

Mindfulness is billed as not just cool and true but useful -

No doubt many distinct mechanisms are involved - the regulation of attention and behaviour, increased body awareness, inhibition of negative emotions, reframing of experience, changes in your view of the 'self', and so forth - and each of these will have their own neurophysiological basis.

Well, I do love self-regulation!

The following argument isn't explicitly stated by Harris, but I find it helpful as an existence-proof for the usefulness of nonessentialism:

1. We are happy and perform well when we're in 'flow' states.
2. Flow states involve "losing" yourself in a task, in a concrete, unhesitating sequence of perceptions and actions.
3. Therefore losing yourself can be good and helpful.

Also

1. We do not directly apprehend the external world; we know it through sense-data plus massive computational modelling tricks in the brain.
2. We know that the brain computes the wrong thing sometimes. (Cognitive biases, optical illusions, top-down processing, hallucinations.)
3. So, if such a thing is possible, it *could* be helpful to attend to sense-data more closely, to spot auto brain errors. *Maybe* more than fleeting sensory illusions too.

While I don't have a very clear philosophy of mind, I know I'm not a direct realist or substance dualist or identity essentialist, so I've no philosophical objections to breaking down the Self, either. Allons-y.

## **Does this stuff work?**

Maybe. For the most important part, mental health, there is a consensus amongst positive and clinical psychologists in favour,  $d=0.3$  or so - but unfortunately this means less than it should. It probably works on average for stress reduction - at least as much as taking a nap does, or valium, or sitting still and breathing deeply for a while. On the other end, it is definitely not the source of brain-juice-drinking power. Somewhere between these two limits we drift, deciding whether to spend time on it.

(Note also that there are likely to be types of people who are harmed by contemplation and self-negation.)

## **Is it worth it?**

It's an expensive project: it costs me part of my most wilful and focussed hours, maybe 3% of all my waking hours, to be spent, if I am serious, for the rest of my life.

Even if I accept that mindfulness is a source of value, there's presumably still a tradeoff against clearer, quicker, more public sources: doing science or kindnesses or pleasures. 10 days spent in myself is 10 days not learning, not exercising, not enjoying, not helping, not meeting, in solitary. (And even on the contemplative axis it competes with Stoicism, with yoga, with writing, with psychedelics.)

It is sometimes claimed that it will increase my focus and so pay off in those narrow terms. But I'd be surprised if the effect was strong enough to overcome the high time investment.

Some contemplatives freely admit that the cost is very high: some contemplatives are not just salesmen. I met someone who claimed to be capital-e-enlightened. (He was otherwise articulate and modest.) He said it took 6 years' work, at many hours a week. I asked him if he could say how valuable it is in other terms - 'What else has been as good?' He said: a decade of intense psychotherapy, or two philosophy degrees.

(One ancient text teases us by setting 'seven years' as the required period, but in true troll-Buddhist style it then slowly walks back this helpful definite statement.)

I was looking forward to writing a gotcha here, but Harris (and thousands of years of arhats and yogis) pre-empted me:

...the deepest goal of spirituality is freedom from the illusion of the self -[but] to seek such freedom, as though it were a future state to be attained through effort, is to reinforce the chains of one's apparent bondage in each moment.

One [solution] is to simply ignore the paradox and adopt various techniques of meditation in the hope that a breakthrough will occur. Some people appear to succeed at this, but many fail... Goal-oriented modes of practice have the

virtue of being easily taught, because a person can begin them without having had any fundamental insight...

... The other traditional response is... to concede that all efforts are doomed, because the urge to attain self-transcendence or any other mystical experience is a symptom of the very disease we want to cure. There is nothing to do but give up the search.

I'm not actually worried by this, because I suspect the full-Buddhist anti-striving thing is unnecessary and... undesirable.

## Grand doubt about grand doubt

Why should an evolved creature have the power to inspect its own sense-data? If we are constantly distracting ourselves with reified thoughts, what evolutionary role did this play? At the top of this review is Harris' droll diss about people deluding themselves into thinking they are meditating - but how can we know that we, or anyone, is not deluded? (Brain scans of inhibited medial PFCs are interesting but merely suggestive.)

This is more of a brain dump than a review: most of the above isn't directly from Harris, I'm riffing off better rational reconstructions of this ancient one-weird-trick. His chapter warning of the history of appalling abuse by gurus and yogis is a public service and I'd be happy to see it in every self-help book.

Some aficionados are a bit snobby about Harris and his app, just as he is aggressive about the religious and cultish sides. I suppose the great benefit of Harris *is* abrasiveness: this is the only way to reach a certain large demographic - the 'epistemic rationalist', the Skeptic, the Freethinker, the parachute RCT wanter. Harris has so much credibility as a rational thug that he can bring mindfulness to its most distant, conceptualising, recalcitrant population. I am open to the idea that this is a good thing.

*See also my thoughts on ways introspection fails.*

<h3>Why listen to me on this topic?</h3>

<div>

<i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect

<ol>

- <li>immersion in the field and/or good priors for what makes for an extraordinary book
- <li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes to write a book
- <li>incredible amounts of argument-checking, which doesn't need domain knowledge

</ol>

I always try to do (3) but surely often fail.</i> <br><br><br>

In this case, don't trust me much. I am no mind scientist; nor have I personally experienced any of the above.

# Yoga hypothesis dump

Gavin

2019-04-13

{% include yoga/links.md %}

Mindfulness is trendy in my social circle, as part of a wider attempt to psychologise and appropriate the good bits from religion (gatherings, singing, self-transcendence, gratitude, existential hope, annual rituals). As a Brit - as a proudly and wilfully fettered person - I am suspicious of all this grinning and trusting and spirit. 2

To my surprise I like yoga. (The hardish kind, ‘ashtanga’.) I never regret going, and usually emerge with a clear head and a warm glow. Why?

At some point I’d like to do a proper literature review on the purported outcomes of yoga vs meditation and so on (with a suitable correction for how bad and inflated medical/psychological research is) but for now I just want to list hypotheses (in the manner of this great post by Katja Grace). 1

1. Just endorphins. It’s resistance training of a kind, so it produces the usual exercise high.
2. Stretching: I don’t stretch outside class. Maybe it’s a *de facto* massage.
3. Deep breathing.
  - Brute oxygenation. Maybe I don’t breathe enough in everyday situations.
  - Most mindfulness gives focussing on your breath massive significance, as a way to occupy your stream of consciousness and sort of stop thinking (or stop identifying with your stream). As a break from thinking this seems plausibly good, but of course some strains of the source religion take it further and treat thought and explanation as an enemy. See hypothesis #6.
4. Giving up agency for an hour: Maybe it’s pleasant to do as you’re told, to not make active decisions (in moderation!).
5. Focussing on concrete things. For me, I think all of the supposedly therapeutic effect of not thinking comes from having to focus on moving carefully, from being actively distracted from my flywheel mind.
6. Corpse mode: At the end of a session you completely let go for 5 whole mins. It feels fantastic (though it doesn’t work without the preceding hour of strain). All exercise would probably produce great effects if it ended in

this way. (That's the literal translation, incidentally.)

7. Group psychology: Most people like doing things in a group.
  - There's an aesthetic angle to doing things in unison, cf. choir singing.
8. Proprioception fun: your eyes are closed for much of the time. Is it a pleasant challenge to move one's body without visual feedback and balancing?
9. Thought of being toned: Usual optimism produced by physical activity.
10. Bare feet: It's pleasant to take off my shoes.
11. Silliness: I quite like silly things, and wobbling in Warrior III is silly.
12. Mystic decorations: Maybe there's a guilty pleasure in listening to woo for once.
13. Inversion: Is there something about being upside down?
14. The company of women. Gender split in my class is 9:1. This is a novelty, doesn't happen in any other part of my life. (Except choir.)
15. Improved concentration. Insight or apparent insight into your stream of consciousness, how perception is really a real-time model of your surroundings.
16. Improved body awareness. Awareness of being an animal

Most of these suggest their own tests naturally. You test (5) by doing your own thing, (8) by doing it alone, etc. The null hypothesis is that it's just (1), which you test by sitting still and trying not to think of anything. (I did. It wasn't nice.)

Even after we find out the active ingredients of the pleasantness, we still don't know that yoga improves e.g. cognition or mental health or productivity, except insofar as feeling good is constitutive of those. (Which it is to some degree.)

What else?

{% include yoga/foots.html %}

# ‘Scarcity’ (2013) by Shafir & Mullainathan

Gavin

2019-05-02

```
{% assign lux = "https://en.wikipedia.org/wiki/Luxury_goods" %} {% assign zvi = "https://thezvi.wordpress.com/2017/09/30/slack/" %} {% assign chok = "https://en.wikipedia.org/wiki/Choke_(sports)" %} {% assign moral = "https://scholar.harvard.edu/files/bfriedman/files/the_moral_consequences_of_economic_growth_0.pdf" %} {% assign z = "http://www.nonzero.org/app1.htm" %} {% assign jung = "http://www.sebastianjunger.com/tribe-by-sebastian-junger" %} {% assign lil = "https://en.wikipedia.org/wiki/Little's_law" %} {% assign hospit = "https://www.strategy-business.com/article/00229?gko=a073f" %} {% assign iron = "https://www.gwern.net/docs/sociology/1987-rossi" %} {% assign slack = "https://www.lesswrong.com/posts/yLLkWMDbC9ZNKbjDG/slack" %}
```

Economics bills itself as ‘the study of decisions under scarcity’, but a lot of it is actually about excess: luxury substitution, savings rates, futures markets, conspicuous consumption... But even these theories are about constrained optimisation, just with looser constraints.

That’s the material side. The psychological side of painful scarcity - the panic, mental narrowing, and sense of doom - was completely absent from my economics classes, but without it you can’t really understand poverty, and can’t value economic growth as the life-saving, *mind-saving* thing it has been.

## Reasons scarcity is bad:

1. Lower consumption is often less good (and sometimes maximally bad, if we count emergency medicine as a consumer good).
2. Less freedom (fewer choices)
3. Anxiety (emotional penalty)
4. Cognitive penalty (bandwidth wasted on worrying about)
5. Excessive focus on the present compromises planning for the future (“tunnelling”)
6. Have to spend more time on careful allocation (“juggling”)
7. Excess self-consciousness means worse performance (“choking”)
8. Can poison social interaction by encouraging zero-sum thinking and so wasteful conflict.<sup>1</sup>

9. It recurses: Mistakes lead to real sacrifice (debt; traps; no slack means penalties bite, further reducing slack). Scarcity causes more scarcity by screwing with your planning and implementation skills.

Economics only really handles costs (1) and (2). Psychology at its best handles (3-7). (9) is the author's new contribution, I think: cognitive economics. The study of decisions under scarcity - but now the internal view.

Without some spare resources it's impossible to be free, to be generous, to relax. That's obvious. Less obvious is that without slack you can't even think straight: there's a "bandwidth tax" on the poor, reducing their effective intelligence and willpower by - apparently - an entire standard deviation. Most of the experiments cited in this are about money scarcity, but their ingenious move is to generalise to all of us, to all conditions where a person lacks some instinctively (evolutionarily) key resource: e.g. money, time, calories, or friends. As well as being a plausible and exciting theoretical synthesis, this makes the book more evocative for rich-world readers:

We have used the psychology of scarcity to create an empathy bridge. We have used experience with one form of scarcity (say, time) to connect to another form (money). Having known what it's like to badly need a little more time, we might start to imagine what it's like to desperately need a little more money or even more friends. We used this bridge to draw a connection between a busy manager fretting about insufficient time before a deadline and a person short on cash fretting about insufficient funds to pay rent. 2

Exciting! I've been reading development economics and behavioural science for years, and I still got a lot of new results and a gosh-darnit gears-level Practical Theory of Mind.

They compress all the constructs and determinants of their real theory into a simplified idea, "bandwidth". This is a shorthand for working memory & fluid intelligence & attention span & decision consistency & persistence & executive control & long-term planning inclination. They admit at the start it's a compression, so that's fine.

With compromised bandwidth, we are more likely to give in to our impulses, more likely to cave in to temptations. With little slack, we have less room to fail. With compromised bandwidth, we are more likely to fail.

Lesson: To optimise your life, you can't 'optimise' too hard, in the sense of pushing right up against your budgets. This idea is not new; a different book would cross disciplines and tie this to queuing theory and distributed computing, trying to find general theoretical truths about systems. (What's the maximum sustainable load for a server? For a life?) Excess capacity, 'slack', is short-run inefficiency and long-term shock-tolerance and thus true efficiency. The point seems to apply to servers, hospitals, and a single human life viewed from inside.

This also adds to Taleb's critique of naive finance, encouraging 'risk-sensitive optimisation' (death-sensitive). They extend bounded rationality from a computation

and search budget to limited attention and willpower.

The book's big philosophical question is the old Essence vs Context chestnut ("the poor are worse parents, drivers, borrowers" vs "given these constraints, people are worse parents, drivers, borrowers"). But it's a new twist: as well as causing permanent developmental deficiencies, poverty levies temporary mental costs:

This shortfall is not of the standard physiological variety, having to do with a lack of nutrition or stress from early childhood hindering brain development. Nor is bandwidth permanently compromised by poverty. It is the present-day cognitive load of making ends meet: when income rises, so, too, does cognitive capacity. The bandwidth of the farmers was restored as soon as crop payments were received. Poverty at its very core taxes bandwidth and diminishes capacity.

This surprises me: I generally accept that people are hard to change, that engineered context is relatively weak. But then all attempts at self-improvement are a denial of essentialism about something, and I'm well into those.

To explain why the poor borrow excessively, we do not need to appeal to a lack of financial education, the avarice of predatory lenders, or an oversized tendency for self-indulgence. To explain why the busy put off things and fall behind, we do not need to appeal to weak self-control, deficient understanding, or a lack of time-management skills. Instead, borrowing is a simple consequence of tunneling.

They don't sugarcoat it: they accept the massive body of evidence on how burdened the poor are, on dozens of axes. And they note that just giving them cash rarely solves the problem because this doesn't change the logic enough.

The poor stay poor, the lonely stay lonely, the busy stay busy, and diets fail.

One big gripe: They use the word "scarcity" for both a physical shortage (i.e. the normal economic sense) and for this special psychological burden. (Not having, and having your mind captured by not having.) This needs two words; it muddies their thesis.

They've persuaded me that late fines are an extremely regressive tax. I'm open to the view that reducing poor people's options is sometimes best for them (e.g. if they are "hurt by the ability to borrow [at extortionate rates]" because it prevents them smoothing their income in a credit cycle). I agree that bandwidth is the deepest kind of human capital.

Their treatment of the mental costs of education is extremely important, given NGOs' blithe promotion of education over all else. (And it's a further argument for unconditional cash transfers.)

To capitalize on a bonus payment for a child's medical checkup, a parent must set up the appointment, remember to keep it, find the time to get there and back, and coerce the child to go (no child likes the doctor!). Each of these steps requires some bandwidth. And this is just one behavior. Conditional

cash transfer programs seek to encourage dozens, if not hundreds, of these good behaviors. Just understanding those incentives and making the necessary trade-offs—deciding which are worth it for you and which are not, and when—requires bandwidth. We never ask, Is this how we want poor people to use their bandwidth? We never factor in this cost in deciding which behaviors are most worth promoting. When we design poverty programs, we recognize that the poor are short on cash, so we are careful to conserve on that. But we do not think of bandwidth as being scarce as well. Nowhere is this clearer than in our impulse to educate.

I'm a keen and cynical student of social research, and but I only recognised one spurious result in the whole book. (ego depletion, p.107 - and that only in a hypothetical tangent.) They did a pretty convincing within-subjects study on sugar farmers before and after harvest income, which nails down the effect as far as I can see.

Only not five stars because we can't give any social science book five stars until it is 20 years old and more severely scrutinised than this.

```
<h3>Risks</h3>
<div>
    The <a href="{{slack}}>Slack</a> doctrine advocates not trying exactly as hard as you do
</div>

<h3>Premortem for unified scarcity theory</h3>
<div>
    If in 10 years we find that the theory above is in fact not a good one, what will the world do?
    * Confounders in the sugar experiments?<br>
    * Strong genetic predisposition to tunnelling.<br>
</div>

{% include scarc/disclaimer.md %}
```

## Scarce evidence?

Camerer 2018 replication and meta-analysis wiped out Then O'Donnell 2021's bundle of 20 replications

Their own rerun found that one of the effect didn't replicate, but "scarcity itself leads to overborrowing" did.

Although see Junger for positive social effects of acute, temporary shared adversity.

It's common to mock people for claiming that they are "time-poor", since the speaker will generally have a high income and will have chosen their situation, while the poor involuntarily suffer both money and time poverty. 'Scarcity' implies that you should ease up on them, since there are serious quality of life

losses from doing too much to keep up your mortgage and your au pair and your fitness and all that.

# Maths at the Open University

Gavin

2019-03-24

{% include ou/links.md %}

Anon: “What would you give, to be two standard deviations better at math?”  
Scott Alexander: “Ten years of life.”

Sometime in 2012, I realised that you can’t do without maths for general impersonal truth-seeking - and that I didn’t know enough of it to do science, or even real data analysis. Nor did I have the confidence to self-study. What to do?

Maths is probably the best subject to learn online, because set problems can *always* be cracked with sufficient thought, and because learning it *can’t be done* without lots of independent thought and silent focus anyway. 1 And Britain has been doing cheap, high-class MOOCs for 50 years, in the form of the giant public Open University.

6 years later, and I’ve got a BSc (Hons) in Maths and Stats, working a full-time job throughout. I know something now. It was good! But it probably wouldn’t be for you, if you’re not strange in the particular ways I am.

To see if it’s good in general, better than my emoting is checking the graduation rate: how often do people see enough value in it / get sufficient help to finish the course? The median completion rate for MOOCs is about 4% (edX); the OU is about 14% for all courses. 2

{% include ou/functions.html %} {% include ou/syllabus.html %}

---

## Benefits

- *Absolutely maximal flexibility.* You *can* do a full degree in 2 years if you’re crazy, or in 16 years if a lot of life happens to you. (They estimate 16 hours a week for part-time study, but I managed with about half that.) There’s a start date every 6 months. OU degrees are even available to sailors on nuclear submarines submersed for months at a time. No lectures - good riddance. Most tutorials are streamed and recorded. The *only* physical requirement is going to an exam centre one week once a year.

- *Structure and tempo.* I found the deadlines and personal tutoring incredibly helpful, relative to getting a textbook and trying to summon willpower. Much better than other MOOCs I've done, too, and not much less motivating than my face-to-face degree.
- *Personal tutor.* Each course has a tutor who you can write to as much as you like, and who respond within a day. Most tutors give you their home phone number - which I never used, but which gives you an idea of the service ethic. The tutors are mostly maths PhDs or veteran longbeards. Once you know LaTeX emailing precise questions becomes viable.
- *Excellent course materials.* These are mailed to you and are also available online. They're high quality and totally self-contained - which is a mixed blessing, since I didn't learn how to handle real maths references text (with their masses of irrelevant results and sadistic 'exercises left for the reader'. I will have to learn this for grad school.
- *Zero entry requirements.* "The university of the second chance": Everybody gets in, and there's a few competence streams to prevent terror/boredom. There's an optional high-school-level course to give you the really basic building blocks. For the highly driven, it's an alternative to school without the brakes: a few kids have speed-run it by the age of 15.
- *Cheap.* OK, so on this I got lucky.

The total cost for an Honours degree (with a Scottish address): £6,048. Total cost (England or international): £18,072.

NB: About 1 in 10 students get fees paid by their employer - it's so cheap, it buys your loyalty, and they can write it off. I got the whole thing paid this way, plus a bunch of study leave.

- *Breadth.* I've got a rough idea of large parts of pure mathematics, even though I took every statistics elective I could. I won't pretend this is more than me being able to learn any subfield now.
- *Time to marinate.* To me, taking twice as long is much better for learning. Lots more time for intuitions to be built, for shower-thought epiphanies, for the pieces to get joined up. This is also a serious test of the spaced recognition technique - I refreshed calculus once a year for six years. (This is a strict positive, despite using up more of your life, because you can do it quicker if you like.)
- *Beautiful, fixable typesetting.* When you're starting out in maths, you constantly make mistakes. (Later, mistakes are only very very frequent.) If you're writing by hand, this leads to hours of wasted effort rewriting fixed proofs. You're taught LaTeX in the second or third course, and from then on all your homework submissions can be in that.
- *Automated drudgery.* Later courses let you delegate lots of the rote work

(like inverting bloody matrices) to computer algebra systems like wxMaxima, trusting you to know what you're doing.

- *Open assessment metrics.* They post the pass rates and top-marks rates for each course. Decades of past papers online too.
- *Prep camp.* There's a student association for OU maths, the M500 Society. They run a cheap annual exam prep camp in a giant hotel conference place in Milton Keynes. It's surprisingly good!
- *Zero group work.* If you're pathologically independent, like me, then this is a large plus. For most people, it is demotivating and low in meaning.

## Problems

- *Distance means dropout risk.* The graduation rate is much better than the average MOOC, but still way below traditional unis. Most of this gap is probably because the OU is so much less selective than the face-to-face unis; so despite appearances the gap is less a bug than a feature. (The remainder of the gap is probably mediated by lack of social interaction and meaning-making.)
- *Not especially deep.* You graze quite widely over geometry, number theory, calculus, diffeqs, first-year physics, combinatorics. As a result, you're regularly returning to elementary matters - so my second year courses were the first time I felt fully challenged. The only thing I covered in any depth was probability theory and stochastic processes, but that's because of my choices.
- *No undergraduate research.* No option for a maths dissertation, which is great if you've no ambitions in the matter.
- *No continuity of teachers.* This mostly scuppers your chances of getting a single strong academic reference (instead there's a centralised bundle of comments from past tutors).
- *Not especially prestigious.* The completely unselective start of the pipeline isn't as bad for the degree's signalling as you'd think, because you need to be pretty strong to make it out the end. You basically lose the entire bottom eight deciles. (One third of graduates in my course get a First, which isn't so easy - takes 85%+ on all final courses.) Anyway it hasn't stopped me getting into a decent grad programme (after doing a bunch of additional side projects).
- *One nonmaths elective.* It's compulsory to take one course outside your major - luckily the Linux / Windows networking one was useful.
- *Bad philosophy of science.* The stats courses are stubbornly crap-frequentist, and require you to parrot false or misleading statements (" $p >$

*0.05, therefore...")* to get full marks. (This problem is far from unique to the OU though.)

- *Crap proprietary software* (MathCAD, Minitab, GenStat, SPSS). The stats courses demand that you install various meh packages. Licences are included in the fee, but it's still a wasted opportunity to learn superior and future-proofed data science tools. I did most of the exercises in SciPy anyway, and only lost a couple points to pedantic markers.
- *Handwritten exams*. I never write with a pen anymore, so I had to spend a couple of weeks building up hand muscles before exams. It's kind of painful.

## Bottom line

On the spectrum between “buy a textbook and sweat it out alone” and “attend 20 hours of compulsory lectures, do 20 hours of compulsory exercises - and spend all your time with people doing the same”, it’s closer to the former. But this was no bad thing, for me.

It doesn’t develop your research skills very much - a *lot* of the homework exercises involve spotting the right algorithm to use, out of a small number of given algos, then turning the crank. (Though I *occasionally* came up with my own method - e.g. using the fundamental theorem of algebra to terminate a root-finder - and got full marks.) Proof is underemphasised, relative to full university treatments.

The full £18k sticker price probably isn’t worth it unless you have really hard constraints on your geography or time. If you can get subsidised - which is pretty easy - and if you’re an introvert, it’s great.

```
{% include ou/foots.html %}
```

# The Tacit Analytic Metaphilosophy

Gavin

2019-03-08

The duration and depth of disagreement among philosophers - 2000 years, on many, many things - is funny. Why are we surprised by this failure?

I think we care because a particular background metaphilosophy is frustrated by persistent disagreement. Call the following the *Tacit Analytic Metaphilosophy*: conceptual realism: concepts have one true definition.

1. *Metaphilosophical realism*: (Suitably sharpened) philosophical questions have one objective answer. 1
2. *Metaphilosophical naturalism*: “Philosophy is just one part of the empirical attempt to understand Nature.” 2
3. *Metaphilosophical optimism*: The effort to obtain answers to these questions is both possible and important.
4. *Metaphilosophical idealism*: The context of discovery of philosophy is not relevant to its context of justification.

And call the denial of (1) *pluralism* or *relativism*; the denial of (2) *anti-realism* or *humanism*; the denial of (3) *quietism*; and the denial of (4) historicism.

There's a clear tension between (2) and (4): how can it be the case that philosophy is just another natural phenomenon, *and* that it (or its justification) float free of particular causal histories?

The sensible reply is that (4) is a normative claim: *of course* all kinds of cognitive bias and personal idiosyncracy are powerful factors in the development of a philosophical view, but it's irrational and counterproductive to focus on that, because there seems to be nothing to say about such factors - certainly nothing polite.

How common is this view? The PhilPapers survey (of mainly Anglophone philosophers) find

# Self-experimentation

Gavin

2019-06-17

factorial analysis of supplements. No external validity sought.

Caffeine sends some people to sleep.

Index 1. Books per month 2. #pomodoros 3. Self-rated productivity 4. Sleep quality 5. Fluid cognition RAPM DNB backward digit span Cambridge Brain Sciences?

Need

- EEG (when sleep on side)

## Chems

Modafinil Nicotine Flow Theanine Caffeine

# Data fumes

Gavin

2019-03-08

Nick Bostrom mentions in passing the idea of “data fumes”: that, when a field is new and priors weak, an early influx of low-quality data can throw it off the trail for decades.

This might be due to anchoring, but also the natural worship of founders.

- Wunderlich on basal body temperature, 98.6 F. Formidable dataset - 1 million observations in 1860! Only discovered by testing the thermometers he used, miscalibrated by 3 F. 150 years later.

# Rebellion, and rebellion against rebellion

Gavin

2019-03-26

```
{% assign int = "https://en.wikipedia.org/wiki/2011_military_intervention_in.Libya"
%} {% assign sing = "https://www.philosophyexperiments.com/singer/" %} {% assign fair = "http://www.michaeldello.com/fair-trade-worse-free-trade/" %} {% assign rec = "https://www.nytimes.com/2015/10/04/opinion/sunday/the-reign-of-recycling.html" %} {% assign lib1 = "https://www.foreignaffairs.com/articles/libya/2019-02-18/obamas-libya-debacle" %} {% assign lib2 = "https://foreignpolicy.com/2018/06/22/the-west-is-letting-libya-tear-itself-apart/" %} {% assign second = "https://en.wikipedia.org/wiki/Libyan_Civil_War" %} {% assign rape = "https://www.independent.co.uk/news/world/africa/amnesty-questions-claim-that-gaddafi-ordered-rape-as-weapon-of-war-2302037.html" %} {% assign feb = "https://www.nytimes.com/2011/05/05/world/africa/05nations.html" %} {% assign islamist = "https://en.wikipedia.org/wiki/Justice_and_Construction_Party" %} {% assign reb = "https://en.wikipedia.org/wiki/Libya_Revolutionaries_Operations_Room" %} {% assign china = "https://www.technologyreview.com/the-download/612577/the-us-has-blamed-chinese-state-hackers-for-the-marriott-hotel-data-breach/" %}
```

The wiser course might often be to do nothing, but it will seldom be without moral cost.

— Clive James

In 2011 my university hosted a debate about the fresh Libya intervention. Alongside the pie-eyed political scientists, some Libyan students were on the panel. They described rapes and massacres, how their families were praying for NATO intervention, how it was the only hope for democracy, how in fact their families were otherwise sure to die.

This was formative for me. I'd protested the 2003 Iraq War (reflexively, ineffectually) when I was in high school. But I'd been coming around to consequentialism, the worldview which forbids no action absolutely. It just made things make sense: suddenly I knew why I felt bad at luxury spending - because the same money could be saving lives. Other things which had seemed so important - recycling, Fair Trade, official foreign aid, metaphysics, poetry - took on ordinary proportions, stopped needling me, fell away. And so on.

I think a blanket rejection of war was the last deontological principle I had. I had a sure and accurate intuition of the horror of intervention. But here, unavoidable, was somebody telling me the horror of nonintervention.

Things got even more dramatic: The audience was packed with Quakers. They believed that nonviolent resistance is a simple and universal method for preventing violence. They were dogmatic, opposing even the no-fly zone; they didn't answer the questions people put to them, about the unarmed protestors killed; they were inarticulate and petulant, criticising the Transitional Council rebels for taking up arms, and forgetting to criticise Gaddafi at all. (In fact they almost defended him - in that particular dodgy New Left way - for his anti-imperialism.)

And yet they were completely correct about Libya, which 8 years later is still at war:

For the ninth time since 2011, rival Libyan factions are slugging it out to control the country's strategic "oil crescent," a coastal strip which begins 100 miles south of Benghazi and arcs westward 250 miles toward Sirte.

Libya has not only failed to evolve into a democracy; it has devolved into a failed state. Violent deaths and other human rights abuses have increased severalfold. Rather than helping the United States combat terrorism, as Gaddafi did during his last decade in power, Libya now serves as a safe haven for militias affiliated with both al Qaeda and the Islamic State of Iraq and al-Sham (ISIS). The Libya intervention has harmed other U.S. interests as well: undermining nuclear nonproliferation, chilling Russian cooperation at the UN, and fueling Syria's civil war. Despite what defenders of the mission claim, there was a better policy available — not intervening at all, because peaceful Libyan civilians were not actually being targeted.

The pacifists were right, even though they're a stopped clock. Amnesty didn't find any serious evidence of rape as a tactic. After the February killings of unarmed protestors, civilians don't seem to have been intentionally targeted by Gaddafi's forces. 1 (They were busy.) The rebels included plenty of horrible authoritarians, as revolutionary cadres are wont to do.

I don't know whether the Libyans on the panel were lying or misinformed, propagandists or victims of the same righteous fog of war that caught out Juppé, Cameron and Obama.

The lesson is twofold: war can be justified and almost never is. Also: disinformation, which has always been war's companion, makes a mockery of journalism and policy, of straightforward evidence collation - and it'll only get worse now they can target you with specific lies.

---

Don't understand me too quickly. The Quakers were right for the wrong reason.

The consequentialist argument against seemingly good wars is simple: it just almost never works. Your prior should be heavily against it. This time is not

different. And this looks like pacifism most of the time, if an unusually watchful kind.

{% include libya/foots.html %}

# Is education worth it for society?

Gavin

2017-07-02

{% include private-and-social/links.md %}

I got a lot of fully-subsidised education: more than 20 years' worth. In educated circles this is seen as an unalloyed good; I am thought to have benefited both myself and society 4. But I find myself seriously concerned that I actually wronged people with the latter 10. Say there are four kinds of benefits from education.

{% include ed/two.html %}

Wonderful things! But if the social ones aren't larger than the social cost, then state education will tend to be taking from society and giving to those who happen to be above-average in nerdiness. 2

Is education a good deal overall, *including for people who don't get it*? I can't actually resolve this question in less than a book. The algorithm is

1. enumerate the (confusing, mixed, methodologically flawed evidence for) benefits and costs
2. put them on a common scale
3. take the ratio

The following is just part of step 1.

## How to think about education's social benefits

We should distinguish private returns (pay, increased confidence, increased knowledge, increased social capital for you) from social returns (productivity, political contributions, cultural reproduction if you like). The former are good, incredibly good, but not a matter for government policy insofar as they include zero-sum benefits, and if there are better ways to spend public funds.

### Productivity

Are educated people more productive? Yes. But did their education cause this? To some extent maybe!

The reason to pay particular attention to the economic side of the social return is not that money is the most important thing, but because anything that doesn't give net economic returns can't be kept up without trading off against something else, like infrastructure, or social care, or life-giving research, or (let me dream) the fate of the world.

## R&D

You might note that academics produce a large proportion of all inventions and new ideas. This too is confounded: science was *more* productive when university intakes were 1% of current levels. And the relationship between basic science and technology is less straightforward than it seems.

## Intelligence

Two of the most careful psychologists I know came out with an astonishing result: that education actually improves cognitive abilities, perhaps 3 points per marginal year. (Clearly this wouldn't scale indefinitely, but even at normal 10 year levels it's a remarkable effect.) And it was a n=600,000 meta-analysis of 142 analyses.

If this doesn't astonish you, then you haven't been paying attention to just how hard it is to raise intelligence (except by correcting malnutrition or severe pollution).

If this estimate is on the correct order of magnitude, then while there are massive private benefits to this effect, the social gains from more capable citizens will be huge.

## Noneconomic gains

The humanist response is that educating your citizens produces huge noneconomic public goods, like critical thought or voluntarism or political purpose or empathy or taste or cultural continuity ("pass it on!").

The private noneconomic return is enormous, larger than the huge private economic return, for some people. e.g. 4 years of relative freedom, away from home, surrounded by bright horny people can be very good for your later worldview, life goals, and mental health. You get space to build yourself new. Or if not build, then to locate yourself in culture, philosophy, and personality space.

More grandly, you can see education as a compiler: you take a young person and a curriculum, and you output a young person with a better model of the world.

PG:

Reading and experience train your model of the world. And even if you forget the experience or what you read, its effect on your model

of the world persists. Your mind is like a compiled program you've lost the source of. It works, but you don't know why.

This totally answers the correct charge that people forget almost everything about high school and their degrees unless constantly using that knowledge.

Or:

it's important to make kids learn specific facts, but not so important that they remember them; teaching someone (eg) Civil War history is "training" a "predictive model" of the Civil War, war in general, and history in general which will survive and remain useful even after the specific facts and battles are long forgotten. I think this is the strongest defense of modern education, given that we do spend lots of time teaching kids things they will definitely forget. But how would you test it?

- Escape from abusive home / a single ideology
- Makes you savvy, imparts a specific set of cultural skills, such that you can get hired and mingle well in the productive sectors. I don't know whether to call these skills productive themselves.

But again, what matters in policy terms is the relative size of social gain and social cost.

## Knowledge

*this section is incomplete*

This is quite easy to check: how much do people remember from uni, for how much of their lives?

- Doctors have forgotten all of their basic science training 5 years out

(I remember being scandalised by some of my peers selling all their textbooks *as soon as* they graduated. But clearly they knew more about social reality than me.)

Cultural continuity - preserving the knowledge and ideas of past generations - depends on a mixture of education and autodidacts. I don't know what the value of preserving a tradition of Hegelianism or Canadian Irish studies is.

But people know this isn't the real reason for education, because they instead emphasise metacognition, "learning how to think":

## Critical thinking

There's a small, diminishing, and temporary effect on critical thinking. (Humanities degrees do not stand out, incidentally.)

## **Virtue**

*this section is incomplete*

This is in the same vein as the old “reading novels makes you empathetic” research programme.

I have no idea if this is generally true - I was a critical voluntarist before university, the most empathetic people I know did not go to university, and most of my Arts peers emerged with none of these things - but I can tell you I had a very good time. And this, the self-justifying private fulfilment, gives me reason to worry about society’s end of the bargain. It’d be very convenient if what (bookish and middle-class) people found most personally fulfilling was also the best thing for all.

## **Politics**

*this section is incomplete*

Does it make people engage more with actual politics? (Not just social media talk: volunteering, running.)

Does it make people more tolerant and cosmopolitan?

## **Friends**

Plausible that the shared adversity forges peer groups into something that can last a lifetime. But where’s the evidence? Do the homeschooled have fewer or less intense?

## **What’s the social cost?**

### **Money**

{% include ed/calc.md %}

Are there better ways to spend £90k per person? (Yes: but let’s limit it to UK recipients.)

- personal tutor at PhD level, 3 hours a day for three years.
- infrastructure
- poverty alleviation

Then there is the great radical alternative: giving everyone an independent income.

### **Time**

Primary and secondary education takes at least 15,000 hours of the most curious and vital years *of everyone alive* 5. Billions of hours of fruitless boredom.

Literacy and numeracy are probably worth this on their own, so factor out primary school, for only 6000 hours of confiscated life.

Then there's uni, in two tranches: people who hate it, and people who drop out.

In the US, 45% of the 20 million annual enrolments do not finish. A lot of this is due to ability deficit (measured by remedial class enrollment), besides the obvious financial reasons. Because of the sheepskin effect - part of a degree is not worth much to the job market - and the low social return on completed education, this means billions of dollars, and millions of years of life wasted. Not to mention the unnecessary stress and humiliation of pushing people into it.

### **Suffering**

You probably know someone who was traumatised by their school years. Even if only 5% of pupils suffer this much, it throws a huge shadow over the social benefit. But even boredom, or unfreedom, or being forced to associate with cruel people count. (One suggestive result: closing schools for coronavirus was correlated with a 20% drop in teen anxiety rate.)

People who suffer from uni are rarer, but I've met a few. They are totally ignored in the discourse, in favour of the Ennobling Creation of Citizens or 4 Year Crazy Party memes. (Again, millions of people drop out and may be left worse off than before.)

### **Credential inflation as perverse redistribution**

*this section is incomplete*

Using school as the main signal of employability is terrible for the many people who cannot handle bureaucracy, being told what to do, pointless makework, or authority. Autodidacts are amazing but rare and rarely respected as much as they should be.

Caplan's contention is that the wage premium of degree-holders mostly comes at the expense of non-degree-holders.

(Other costs: student debt distress, bondage.)

### **What's the alternative?**

#### **1. Wealth for all**

The most dramatic counterfactual: the government just giving you the money they would have spent on you, after 20 years of investment returns:

what if the government had taken this figure and invested it in the stock market at the moment of your birth? Today when you graduate college, they remove it from the stock market, put it in a low-risk bond, put a certain percent of the interest from that bond into keeping up with inflation, and hand you the rest

each year as a basic income guarantee. How much would you have? \$15,000 a year, adjusted for interest. We can add the \$5,800 basic income guarantee we could already afford onto that for about \$20,000 a year, for everyone. Black, white, man, woman, employed, unemployed, abled, disabled, rich, poor. Welcome to the real world, it's dangerous to go alone, take this. What, you thought we were going to throw you out to sink or swim in a world where if you die you die in real life? Come on, we're not that cruel. So when we ask whether your education is worth it, we have to compare what you got – an education that puts you one grade level above the uneducated and which has informed 3.3% of you who Euclid is – to what you could have gotten. 20,000 hours of your youth to play, study, learn to play the violin, whatever. And \$20,000 a year, sweat-free.

## **2. Grad tax**

The above could be taken as an argument for fees: “the individual plausibly benefits more than society, so let them cough up a bit”. But substantial fees are pretty much a shitshow, certainly in the high-interest, inexorable, cartelized form that exist England and America, where the prices are uniform and useless. But (if we cannot tear down this credentialist bullshit, as below) then certainly a graduate tax is fully justified.

## **3. Regulating credential pollution**

Rather, we should replace the hegemony of higher education - make it so that young people don't need a degree to get decent jobs, or in fact most jobs (besides doctor and pilot and so on).

In extremis, we could make education a protected category in job interviews. We would rely on actual portfolios, entry tests, and work trials (which are open to all and actually measure the relevant quantities) instead of pompous paper. (Aptitude tests are illegal in some American industries, so you'd have to reverse that first.) This would be a more powerful intervention against inequality than free fees, because it would catch the many smart people who do not fit the conformist, examination form of ‘training’.

It might take something as radical as this to stop students defecting against each other and continuing the ruinous cycle. (Besides making education level a discriminatory question, a full basic income would work, too.)

The problem with equalising the status of graduates and nongraduates is that higher education is feted by absolutely bloody everyone: parents, governments, giant corporations, reptilian economists, frothing radicals, whether anarchists, neoliberals, or Juche cadres. (Everyone except a minority of libertarians.) The uniqueness of its cross-cultural appeal means that it is presently the only way that young people can possibly get 4 years of relative freedom to locate themselves, and to do so surrounded by people from all around the world, and to do so in an atmosphere which rewards many kinds of deviance.

You could maybe do this by funding (voluntary) international service; basically giving working-class kids some gap years, too. The cult of travel is nearly as powerful as the cult of school, after all.

```
<h3>Higher Ponzi</h3>
<div>
    Above I argue that academia might be supporting the lucky at the expense of the unlucky
    <!-- -->
    Expanding the student intake causes credential inflation, which feeds back to expand the
</div>
<!-- -->
<h3>Jock the radical</h3>
<div>
    My great-grandfather, Jock Middleton, left behind an amazing library. He was a farmer
    My granda (who ended up farming the same land) used to grouse about this, 70 years later
</div>

What I'd do differently

<div>
    {% include ed/diff.html %}
</div>
```

## See also

- Chris Olah
- Sam Knoche's skin in the game
- The counterintuitively humane Bryan Caplan
- Alex Danco on alternative academic communication and gatekeeping.
- Linda on PhDs

```
{% include private-and-social/foots.md %}
```

# On inhabiting a giant corporation

Gavin

2019-07-15

{% include js/lazyFrame.html %} {% include corp/links.md %}

You find yourself inside a machine the size of a city. It is slow, powerful, theoretically immortal, and contains thousands of cogs and hard-wired operators and inexplicably sealed bulkheads. It's warm and well-watered. You could die in here.

<a href="{{mechImg}}"><div id="mechImg"></div></a>

Which is to say that I worked a pleasant job at a multinational insurance corporation for 3 years. It wasn't dreary, possibly because I was in the "data science" bit, the bit allowed to do new things without strangulating oversight and backwards-compatibility.

They're good jobs, as jobs go. Extremely flexible hours, challenging nonroutine tasks, unlimited remote work, very good pay per hour, massive amounts of autonomy (relative to managed manual work), friendly smart colleagues. And what Zed Shaw says about programming in nontech companies was true here -

Programming as a profession is only moderately interesting. It can be a good job, but you could make about the same money and be happier running a fast food joint. You're much better off using code as your secret weapon in another profession. People who can code in the world of technology companies are a dime a dozen and get no respect. People who can code in biology, medicine, government, sociology, physics, history, and mathematics are respected and can do amazing things to advance those disciplines.

Many people's relation to their employer is that of a servant in the household of the firm. But service is only tenable if you're aligned with your patron. A lot of people are incapable of being paid to care about things. They cannot settle for indirect fulfilment, indirect passion, indirect goods. I was one once; if you had told me ten years ago that I would happily spend my time in such a place, I would have been horrified: I was a full-on acolyte of Bakan, Klein, Chomsky, who attribute most of the world's ills to corporations (or rather to The System which corporations are thought to control).

It's easy to see the pathologies and harms of corporations. The benefits of these unsympathetic machines is hard to see without data. 2

---

### **1 in 100,000**

In a sense there's nothing new or weird about being in such a large organisation. After all, I'm "in" a loose organisation of 66m people, Britain, and similarly I am 1 in the 508m of the EU. (For now.) 1

Venkatesh Rao thinks *all* firms are dysfunctional, their quality utterly unstable, their size a *measure* of their decay:

organizations don't suffer pathologies; they are intrinsically pathological constructs. Idealized organizations are not perfect. They are perfectly pathological. So while most management literature is about striving relentlessly towards an ideal by executing organization theories completely, this school would recommend that you do the bare minimum organizing to prevent chaos, and then stop... It may be horrible, but like democracy, it is the best you can do. Today, any time an organization grows too brittle, bureaucratic and disconnected from reality, it is simply killed, torn apart and cannibalized, rather than reformed. A Sociopath with an idea recruits just enough Losers to kick off the cycle. As it grows it requires a Clueless layer to turn it into a controlled reaction rather than a runaway explosion. Eventually, as value hits diminishing returns, both the Sociopaths and Losers make their exits, and the Clueless start to dominate. Finally, the hollow brittle shell collapses on itself and anything of value is recycled by the sociopaths according to meta-firm logic.

This cynicism is frightfully exciting and flatters your taste and mine. It fits the great romance of the age, the startup. But how to reconcile it with the economic consensus (cf. Baumol and Cowen) that big firms are good, incredibly good, for the people who use their stuff and the governments who manage to tax them? I suppose we have to infer that the mess inside can't stop humans from doing productive things.

It's not just the number of people that makes an organisation unmanageable. The firm's internal software landscape is a comparably vast overhead. I don't think any single person, or any 20 people, really understands the hundreds of legacy systems our team relied on.

---

### **The Bureaucratic Universe**

David Graeber says a lot of false things. But I still read him because, in between those, he says large true things. For instance, he was the first person I heard

pointing out one of the largest facts about the world: capitalism, communism and the mixed economy are all founded on the same social structure, the bureaucracy. (HR are the obvious private example, being a rigid impersonal force with great legal power over individuals.)

A standard Left metaphor for a corporation is a shark: a highly optimised and optimising creature which destroys all other human value in the pursuit of profit. Anyone who has spent much time in a large corp will laugh and laugh at this depiction, remembering the 5 hour meetings, compulsory useless training weeks, the constant duplication of effort, in short the rigid and unprofitable bureaucracy...

That said, we shouldn't equivocate between state and private bureaucracies, as the anarchists do. State bureaucracy is nastier, more threatening, extracts more.

- I just applied for a council tax reduction, since I'll be unemployed for a while. The council asked me roughly 200 questions, including half a dozen ID numbers, and so took an hour and a half of my life. They asked when my partner last entered the country. They asked how long we've been together. Paperwork is a regressive tax on the stressed and the poor, and it should be disincentivised, treated as paid labour. *No* bureaucratic task at BigCorp took this long, except the initial interviews.
  - When I was (briefly) in the civil service my blog posts were scrutinised quite closely. No one ever gave a shit in BigCorp.
  - To be less petty: consider how many people the state (including the most powerful bureaucracy of all, our military) kills, relative to corporations. (You might try and count say pollution deaths on the corporate side, but that's a mistake - command economies were in general even more polluting.)
- 

### The City and the City

<a href="{{prettyImg}}"><div id="prettyImg"></div></a>

- The Square Mile is a pretty, storied place - though it smells. The low-key stink around the central City is due to its ancient drains having to handle about 1000 times more people than they were built to, and its tall buildings boxing narrow streets, preventing ventilation. A lazy novelist could use this to connote moral corruption, but there you go.
- I wish I could say that being surrounded by beautiful buildings - nigh unto greatness - had a sustained effect on my character or even my mood but I'm afraid it didn't. Aesthetics is a treadmill.
- The classic image of a Cityboy is wrong: real conformist / careerists wear a surprisingly bright blue suit.



<br>

---

### Insurance: the tedious thrill

- I kept a volume of Kafka on my desk at all times.
- An insurer is like a bank, except you can't withdraw from your account unless something terrible happens to you. Then you win 100 times your balance. So it's a backwards-casino bank.
- One of the nice things about insurance is that the companies' interests are much more aligned with yours, the policyholder's, than usual in capitalism. If you become safer (after paying them the premium) they make more money. There are a few products that follow this gradient - flood detection gizmos for your pipes, discounts for doing exercise and quitting smoking, and so on. But not that many. See here for why insurers don't do more risk mitigation.
- Related: The firm had an "emerging risks" research department, quite forward-thinking, doing epidemic modelling and future climate shocks. But the health insurance people concede that pandemic modelling is useless - since the claimants die before they can claim for health. (The life insurance people model it.)
- You pay an insurer to hold your risk for you. There is a sense in which a large house insurer "owns" millions of glass windows, millions of water pipes. (But only about 50% of profits are your payments; the other half come from investing your premiums.)
- Even though it's a financial company, in the financial heart of the world, making half its money off capital gains, it's not the same as the big guns. Some posh products like paid bank accounts have two categories for financial industry companies: "Finance I" (hedgies, quants, some brokers) and "Finance II" (deposit bankers, insurers, analysts). So ordinary people on ordinary London professional pay. I only called myself a 'Cityboy' when I wanted to shock pious refuseniks.
- Two regulatory changes caused an astonishing amount of frantic work: tens of thousands of extra hours. One of these was a stroke of a pen.
- Only 1% of the staff were actuaries, doing the distinctive work of the industry. The business-school notion of a "profit centre" (the part that makes the money) is spurious and nasty, but it's extremely useful if you're a rent-seeker looking for unearned rewards. Or an anthropologist seeking how people in organisations actually think. (The inventor of the notion calls it "*One of the biggest mistakes I have made*".)
- What is The Actuarial Problem? 4

- First, pass each customer through a boolean function of handcrafted rejection rules.
  - If they survive, predict their expected loss. A stochastic model with a heavily skewed non-negative response distribution, with a multiplicative structure.
  - *Pricing constraints*: pro-rata (price per time); add some hand coded loadings and discounts (e.g. “expected loss should increase with the sum insured”); enforce monotonicity to prevent customer outrage; similarly, enforce only small changes on previous pricing for each customer.
  - *Regulatory constraints*: remove all explicit factors relating to protected classes and vet factors for strong correlations with them; model explanation; known relationships between risk and risk factors.
  - A pair of GLMs modelling the “severity” (loss amount conditional on claim) and “frequency” (claim probability) handle all this quite well, but boosted trees are edging them out after thirty years.
  - Break the loss into “perils” (categories of risk like injury, third-party liability, accident). Build one pair of GLMs per peril, and sum the products of these to get the per-customer estimated cost.
  - There’s also fraud and “rate raiding” detection.
  - “If you get a raise every year, they’re not paying you enough.”
  - There is a chasm between the builders and the operators of algorithms / mathematical methods. Library maintainers vs library callers. (It isn’t as simple as academics vs private sector - every functioning company will have a few builders, since no algo is so general that it works well without local knowledge.) Actuaries are mostly operators.
  - How BigCorp handles its own financial risk was pretty impressive: an enormous apparatus for internal retrocession, central holdings, international arbitrage of capital requirements (i.e. holding money in nearby countries to minimise the impact of regulations).
  - You could use an insurer’s ‘loss ratio’ (payouts / revenue) as a fairness estimate. 50% is an even split between the two parties. 90% is more usual in UK personal insurance.
- 

## Outsourcing

About half the team were in India.

The classic case against globalisation has two components

1. *Welfare*: “Western companies hiring in the developing world have unacceptable working conditions.”

2. *Fairness*: “Western companies hiring in the developing world pay much less for the exact same work.”

The first is often true (though even sweatshops are often average by local standards) but wasn’t at BigCorp.

The second is true by definition - the companies simply wouldn’t outsource if it wasn’t. In the case of entry-level actuaries it’s about “4 lakh rupees” (£5k) vs £32k in the UK. I occasionally talked to Indian colleagues about this (over drinks, out of management’s earshot), and they were always pragmatic about it - “better this than no outsourcing and no job”. I don’t know whether they should be less pragmatic.

---

### Abuse of terminology

{% include corp/jargon.html %}

- Every new thing gets called “AI”. Except anything invented before the 90s is not AI, even if it is a statistical learning method like the other things you call AI. GLMs are not AI. RPA is AI.
  - My company’s name was intentionally selected to mean nothing in any language. This is a *great* metonym.
  - About half of job titles were inflated. The most common was reskinning your Actuarial Analyst job as a Data Scientist job. As far as I know no-one was ever called on this, and references rarely corrected it.
  - You may have noticed that in modern business, everything is ‘award-winning’. This is due to the incredible array of trade magazines and their trade awards. A charade of ladders. A veneer.
- 

### Corporate ‘AI’

- The outlook is not good: probably most corporate data science projects fail. The defectors vary a lot, but inflated expectations, legacy system hell, GIGO, and a weak engineering base are usually implicated.
- The stages of data science are
  1. Build pipeline and dashboards (1b. Get important people to pay attention to them)
  2. Build predictions (2b. Get important people to pay attention to them)
  3. Build decision system (3b. Build good decision system; 3c. Actually use decision system)

I think almost every DS department in the world is stuck on 1b. We got to (3) in my tenure, starting from nowhere.

---

### The normalisation of deviance

A bureaucracy has great power to obfuscate and punish obvious infractions, but is much too weak to regulate the larger part of work: minor and gradually escalating deviance. Unwritten rules beat written procedure. One of the nastiest pathologies of teamwork is the “normalisation of deviance”, the tendency for work norms to mutate into lazy and harmful forms via social proof.

If a piece of wrongness goes unchallenged the first couple of times, it becomes invisible, it suddenly looks right because everyone else is doing it. Say you go outside your spec - but then nothing bad happens, so then we go a little further beyond the spec...

(Just one example from BigCorp: no-one knew how to Procure a GPU through the Procurement platform, so we did weeks of deep learning on CPUs. In 2017.)

Foone Turing:

My point with this is not to say “HEY PEOPLE STOP BENDING THE RULES,” exactly. It’s that you have to consider normalization of deviance when designing systems: **How will these rules interact with how people naturally bend the rules?**

Disasters aren’t caused by one small event: it’s an avalanche of problems that we survived up until now until they all happen at once. People don’t automatically know what should be normal, and when new people are onboarded, they can just as easily learn deviant processes that have become normalized as reasonable processes...

Dan Luu:

people get promoted for heroism and putting out fires, not for preventing fires; and people get promoted for shipping features, not for doing critical maintenance work and bug fixing.

To prevent your culture from lulling you into insane behaviour:

1. Pay attention to weak signals
  2. Resist the urge to be unreasonably optimistic
  3. Teach employees how to conduct emotionally uncomfortable conversations
  4. System operators need to feel safe in speaking up
  5. Realize that oversight and monitoring are never-ending
- 

### Competence

“You’re technical, aren’t you?”

“Eh, kinda”  
“What do you do?”  
“I’m a data scientist.”  
“Well then of course you’re technical!”  
“Eh. When I played saxophone I always compared myself to Coltrane and Parker; when I do tech I have in mind Feynman, Tukey, Turing, Gwern.”

---

## Hiring

The goal of the future is full unemployment!

- Arthur C Clarke

I ran my first hiring round here, and a dozen more after that. It’s *incredibly* hard, even when you have objective standards like code quality or ML performance to rely on. Pretty much all easily obtained evidence is a really weak signal about the candidate’s actual performance in the job. I won’t complain about having to do homework for jobs again: turns out it’s scary and hopeless being on the other side of the table too.

I enforced a standardised test on our hiring, with a consistent numerical marking rubric. And we got blinding of applicants put in. We set applicants a basic supervised learning problem. About one in eight answered it adequately. Total obvious plagiarism was very common, maybe one in six. PhDs did no better than Bachelors. Very few had a Github or similar code host, a very cheap way to show me that you’re curious / knowledgeable / whatever.

---

## The deviance of turnover

I was surprised by how much fuckery there was from colleagues at the end of their tenure. People who make it into these places tend to be *very* good at regulating themselves, tend to be agreeable and compliant. In three years I remember exactly one raised voice, and one instance of silent fury.

But turnover was high, 30% per year 3, and this was when you saw normal human deviance. Out of perhaps 30 leavers, we had

- 2 people put on ‘garden leave’ (paid to go away);
- 2 people not really showing up during their notice period;
- 2 people openly watching TV at their desk / playing with MuJoCo completely unrelated to work.

The other notable antisocial moment was the honesty box. A fridge of snacks was installed, with a price list and honesty box for payment. Every week it came in £40 (~50%) under its sale balance, so they took it out.

---

So if many corporations have net positive effects, if the work is ok and the culture friendly, and if you can easily redirect your excessive remuneration to what actually matters - why stop?

In my case it was because I tried out direct do-gooding work and found that I was pretty capable, and that there's a great need for more than money.

It also stomped the importance of environment into me. You *can't* do great work if you don't actually care. It's hard to respect yourself if you don't respect what you're doing. And you assume the form of your colleagues to a shocking degree. This can lead to a slow, subclinical, and ultra-privileged kind of burnout:

I just didn't think my work was very important. I would be very depressed on projects, make slow progress, at times get into a mode where I was much of the time pretending progress simply because I could not bring myself to do the work. I just didn't have the spirit to do it...

Over time I got depressed about this: Do I have a terrible work ethic? Am I really just a bad programmer? A bad person? But these questions were not so verbalized or intellectualized, they were just more like an ambient malaise and a disappointment in where life was

(It's easy to shout Marx bingo when you read this kind of thing, and it's not wrong. But that's Marxism: decent negative critique and no practical positive change. It's hard to see how we could have an unalienated society - without much better technology to act as our drudges, anyway. Certainly no actually-existing socialism managed it, and most seem to have made it worse.)

If, once you're financially secure and ensconced in a house and a family, you have no further ambitions, then these places are as good as it gets.

## See also

{% assign src = "https://srconstantin.wordpress.com/2017/05/09/how-much-work-is-real/" %}

- How Much Work is Real?

{% include corp/foots.html %}

# Big List O' Useful Practices

Gavin

2019-04-25

How we spend our days is, of course, how we spend our lives... A schedule defends from chaos and whim. It is a net for catching days. It is a scaffolding on which a worker can stand and labor with both hands at sections of time. A schedule is a mock-up of reason and order — willed, faked, and so brought into being; it is a peace and a haven set into the wreck of time; it is a lifeboat on which you find yourself, decades later, still living.

— Annie Dillard

{% include process/links.md %}

Life is long but finite. Most of our actions are automatic/intuitive - not really consciously pursuing particular goals. (We don't pay attention to how we allocate our attention.) Most default actions end up not helping us with our goals. So we suffer unnecessarily, or make others suffer.

You have value; you could have more. Few people are the perfect lived expression of their highest values - I think they'd probably be insufferable. (Ayn Rand comes to mind.)

For about a year I've been collecting practical tricks that are supposed to improve you in some way. The problem is that self-help, and its internet successor "lifehacking", are full of nonsense.

These can be as humble and uncontroversial as "always put your keys in the same place when you come home" or as weird as "try out 100 different biochemicals and quantify the results", or as abstract as .

Until recently I didn't really reflect on myself, preferring to charge ahead and just try stuff.

These mostly take the form of overriding habits or instincts with external help. There's something to be said for not ordering yourself around. There's something to be said for using intuition as an information source (about yourself). There's a *lot* to be said for sometimes doing nothing in particular (including meditating, you sneak!). If you want to be an idler, be an idler: it's fun and surprisingly dignified. But if you want to do something big, you could probably do with more structure.

Despite appearances this isn't inimical to creativity or freedom: professional artists are most vocal about their intentional 'process'. [https://en.wikibooks.org/wiki/Introduction\\_to\\_Art/What\\_is\\_Art](https://en.wikibooks.org/wiki/Introduction_to_Art/What_is_Art)  
I've bolded the top 10, the ones that help most.

## Meta

1. Bug list. Can't improve things if you don't know what's wrong. routine; work; hobbies, habits, skills; environment; wastes of time; blind spots; fears and aversions
2. Good sleep. Enables everything else. i.e. No screens after 10pm, no caffeine, eye mask, foam pillow. Matthew Walker's book is scary enough to have actually changed my behaviour despite not telling me huge amounts of new info. It puts the opportunity cost ("but I could read another chapter if I stayed up") on the right scale.
3. Learning.
4. Quantified self. Local validity. Humans are too causally dense to do much more than local inference. This is inefficient but what you gonna do.

## Mental support

### Extended mind

Reduce temporal clutter

Clear your cache (prevent rumination and re-eval)

Fix intentions

Anxiety down

Reassure that everything future important is already written  
personal knowledge throughput

Captured

Goals, Plans, Arguments, Data

SMART

Specific: Narrow your focus. "Lose 30 lbs" > "Get in shape."

Measurable: Know what it will look like and feel like when you've succeeded

Attainable: Keep your goals in-line with what you believe you can really do

Resonant: Look inside to make sure the goals are truly meaningful to you

Time-bound: Have deadlines for each of your goals and stick to them.

All books

Basic sleep, mood, productivity scalars

Job applications

Project ideas

Aphorisms

mine and others'

New words

Large spends

Milestones

Released

- My appearance
- All articles
- All papers
- Search history
- Food spends

Blog

- writing
- Mastodon for squibs
- Self-Presentation

Workflowy

- Todo
  - Day plan, Life plan
  - Ultimate value lodestar
- Bullet journal
- Calendar
- Goals, Plans, Arguments, Data
- Digital inbox / offloading addressable memory
- Mood tracking, sleep tracking
- anti- Betriebsblindheit
- mindless routine habit living

Monthly reports

Remarks

- Goodreads as my intellectual history
- Albums and songs

Aphorisms

- Job odyssey
- Teachers list
- Password manager

Frameworks

- rationality, the very idea
- Sequence vs Cluster
  - <https://blog.givewell.org/2014/06/10/sequence-thinking-vs-cluster-thinking/>

Maths

Self-symbolising

- I am easily sated by socialising, impossible to sate with thought.
- I like attention but rehearse gaffes for years
- Time, money, willpower, intelligence, looks, status
  - Time: more than a peasant or parent, less than a 1%
  - Money: more than average, less than a 1%
  - willpower: more than my peers, less than a high-power
  - Int: all I need
  - Looks: above average
    - good jaw, big blues, tallish, broadish
  - Status: relatively high. Not max privilege but rounded stats

- Glasses
- Vegan
- Left-handed
- Scots? not really

Class: ????

- Elite interests, education, income, vocab, habits, scope
- Some network
- Working-class background
- No family subsidy, but small safety net
  - (room and board in the middle of nowhere)

indicators of attainment

- Goodreads - breadth, being above art, productivity, openness
- ~Degrees - unmentioned so humility
  - countersignalling
- Queal - rationality, productivity, openness
- PredictionBook - rationality, productivity, empiricism
- Pink glasses - secure in sex, openness
- Savings
- No: flat abs, homeownership, car, Instagram,

My culture

- Original
  - family norms
  - NE practices,
  - street scenes,
  - local architecture
  - local cuisine,
  - slang
  - self-deprecation and pessimism

Ranked identities

- Scientist / Empiricist
- Not a joiner
- Scholar
- Writer
- Blogger
- EA
  - Charity nerd / charity snob
- Cyclist
- Data scientist
- Aspiring rationalist
- Philosopher
- Economist
- European
- Scot
- British
- Poet
- Punk

Left-hander  
Atheist  
Working-class  
    parents  
    role models  
    initial aspirations  
    no post-22 financial support  
"Disabled"  
    Raynaud's  
    Mild intestinal damage

Relinquishments  
    Contrarian  
        97: "Never have a girlfriend"  
        01: unendorsed Scottishness  
        hXc  
    Shy punk  
    07: Straight edge (after 4 years)  
    07: Sensitive outsider (after 7 years)  
    Humanist  
        10: Music, poetry (after 5 years)  
        13: Philosophy (apriorist, holist, heterodox, ) (after 5 years)  
        14: Ordinary progressivism (after 10 years)  
    Technician  
        15: poverty (after 8 years)  
        16: animals (after 10 years)  
        17: prioritisation (after 1 year)  
        18: sour grapes anti-academic (after 6 years)  
        18: caffeine (after 4 years)  
    Searcher

Personality continuum:  
    joking: enough  
    scepticism:  
    assertiveness: bit low  
    amount of eye contact: bit low  
    directness: bit much  
    laziness at work: too much  
    laziness at life: bit low

Models  
    Bayesian reasoning  
        assign a probability before an event, receive evidence, update the probability.  
        Your intuition is not worthless;  
        forces us to think probabilistically;  
        allows us to count competing evidence;  
        promotes a nuanced view.

Inside view / outside view  
    Reference class forecasting

To not suffer the planning bias, ignore your particularities and look to similar  
Signalling actions convey information about the actor. Buying an expensive wedding ring conveys wealth.  
Countersignalling: intentionally doing something low status to show that you are not important.  
Anti-signalling: intentionally not signalling, to show that you (think you're) not important.  
Nudge: choice architecture the presentation of choices can have dramatic effects on decisions.  
Trivial Inconveniences Mean We Give Up  
Marginal thinking what are extra resources worth? Less and less.  
System 1 (Automatic) vs System 2 (Deliberate) decisions use two different systems. System 1 is fast and subconscious: 'gut'. We often act on System 1.  
Aliefs / Beliefs automatic tendency / endorsed rational claims  
Comparative vs. absolute advantage positive-sum trade is always possible; people should do the action that they're better at.  
Illusion of transparency we tend to overestimate how much our mental state is known by others. e.g. rhythm of our speech.  
Opportunity cost payoff minus next best payoff. if you don't spend your time in the best way possible, you'll regret it.  
Cognitive biases systematic flaws in how we think.  
We don't look for information that proves us wrong.  
We estimate easily recalled events as more likely than they are.  
We're overconfident. Particularly relative to our level of expertise.  
"Only our enemies are biased."  
"only our friends and colleagues who are biased."  
Heuristics we use rules of thumb to make decisions quickly. notice that we are using them here.  
Counterfactual reasoning what else would have happened in the absence of x?  
e.g. pushing the paramedic out of the way and do the CPR yourself, but you'll die.  
even if you stop the patient from dying, your (counterfactual) impact is still there.  
Doctors wanting to make a big difference will do more good if they travel further.  
Expected value the probability of each outcome multiplied by the value of each outcome.  
50% chance of \$100 is worth \$50, so be willing to invest up to \$50 for a chance to win \$100.  
Time value of money we'd prefer to have \$100 today than tomorrow, or in a year. Thus money today is worth more.  
Money value of time time is a scarce resource, can be used to earn money or do something else of value.  
Even if driving to the other side of town saves you \$10 on groceries, if it takes an hour, it's not worth it.  
Prisoner's dilemma players choosing between cooperating and defecting.  
Co/co = +/+.  
Co/def = +/--,  
def/def = -/-.

equivalent to Tragedy of the commons  
e.g. country not limiting carbon emissions; e.g. athletes doping; e.g. do

Moloch

Revealed preferences  
talk is cheap, our actions reveal more about us than we'd like to believe.  
Polls vs ballots  
Tool for controlling for bs and hypocrisy

Typical mind fallacy  
forgetting Neurodiversity mistaken belief that others have the same mental exp  
'see' a zebra in their mind's eye, some recall the vague idea of a zebra, ot

Fundamental attribution error  
attributing others' actions to their personality, but your actions to your situat

Aumann's agreement theorem  
if two rational agents disagree when they have exactly the same information, one

Bikeshedding  
substituting a hard and important problem for an easy and inconsequential one.

Meme  
a social gene, unit of selection for rituals, behaviours, and ideas.  
Contagion, mutation, selection gradient  
the most useful memes don't necessarily get reproduced.  
there should be few major social movements that don't have "make more like-mind

Algernon's Law  
Intelligence is adaptive  
So we shouldn't expect too many easy nootropics

Social intuitionism  
moral judgements are made predominantly on the basis of intuition followed by ra  
slavery was previously the norm, thought to be acceptable.  
blind faith in intuitions can be harmful and counterproductive, because they can  
It wasn't just that proponents of slavery got the wrong answer: they had - and w

Apophenia  
tendency of humans to see patterns in random noise. e.g. hearing satanic mes

Goodhart's law  
Campbell's law  
what gets measured, gets managed (and then fails to be a good measure).  
If we use GDP as a measure of prosperity of a nation, and there are incentives t

Moral licensing  
doing less good after you feel like you've already done some good. After donati

Chesterton's fence if you don't see a purpose for a fence, don't pull it down  
If you can't see the purpose of a social norm, like whether there is any value t

Peltzman effect  
moral hazard of safety  
taking more risks when you feel more safe.  
When seatbelts were first introduced, motorists actually drove faster and closer

Semmelweis reflex  
not evaluating evidence on the basis of its merits and instead rejecting it beca

Bateman's principle

the most sexually selective sex will be that which has the most costs from rearing

Hawthorne effect  
observer effect; people react differently when they know they are being observed

Bulverism  
attacking the context of discovery rather than the context of justification  
dismissing a claim on the basis of how the opponent got there, rather than a reason  
"but you're just biased!" or "of course you'd believe that, you're scared of the truth"

Schelling point  
a natural point people converge upon independently. participants may not even realize it  
e.g. "no genetic editing of humans, because we don't want inequality"

Local vs global optima  
you might need to make things worse to get to the best possible place. to earn more money

Anthropic principle  
you are given a sleeping pill which will wake you up twice if heads, and once if tails  
Similarly, what is the chance that we're in one of the only universes that is compatible with us?

Arbitrage  
taking advantage of different prices between markets for the same products. Money

Chaos theory  
the present determines the future, but the approximate present does not approximate the future

Ingroup and outgroup psychology  
xenophobia, the left-right political divide. Our circle of concern is probably too narrow

Red Queen hypothesis  
organisms need to constantly evolve to keep up with the offense or defense of their predators

Schelling segregation  
even when groups only have a mild preference to be around others with a similar background

Pareto improvement  
a change that makes at least one person better off, without making anyone worse off

Occam's razor  
among competing hypotheses, the one with the fewest assumptions should be selected

Regression toward the mean  
if you get an extreme result (in a normal distribution) once, additional results are likely to be closer to the mean

Cognitive dissonance  
holding two conflicting beliefs causes us to feel uncomfortable, to reject (at least one)

Coefficient of determination ( $R^2$ )  
how well a model fits or explains data (i.e. how much variance is accounted for)

Affective forecasting  
predicting how happy you will be in the future contingent upon some event or change

Fermi calculation  
sequence of guesses and simple arithmetic to estimate.

Elephant in the brain  
adaptive self-deception

Value of Information competing access needs

Extra

Efficient market hypothesis  
a lot of clever people value approximately the same things as you. Everything is rationalized

<https://conceptually.org/long-list-of-concepts>

Do the hardest thing first  
<https://sapien.co/productivity-hack-1/>

### Checklists

#### Procrastination

Increase expectancy of winning

Expecting to make it

Optimism

Success Spirals

achieve one goal after another

pay attention to your successes

Vicarious Victory

be around successful, positive people

Mental Contrasting

After imagining what you want to achieve, mentally contrast that with learned helplessness

Increase payoff

Flow

increase difficulty until your skill level is stretched

Meaning

connect to something you care about for its own sake, at least through a extrinsic reward for completing it.

Passion

Decrease impulsivity

precommitment

throw away the key

unplug your router

pain

Beeminder

energy

set goals

subgoals, to the one-day task level

Decrease delay

move it forward

#### Prose

#### Publishing

Check for uncited prior art

verbatim plagiarism

philosophical plagiarism: ideas and the relation of ideas

translation plagiarism:

cryptomnesia:

compression plagiarism: distil a lengthy scholarly text into a short one, possibly changing the meaning

#### Stats / ML

impose structure on projects

+ Large life choice criteria

Environment, Mastery, Culture, Communities, Autonomy, Purpose  
Environment

Is the city pleasant to be in?  
Is there good public transport?  
Is there easy access to nature?  
Can I afford to live comfortably?  
Will I have to commute?  
Is the weather good?

Culture

i.e. high density of kinds of people I want to meet  
What kind of people/lifestyles are accepted/respected? (eg SF bipolar at)  
What are the main industries?  
Are there good universities / research groups in the area?

Communities

Do I already know people in the area?  
Is there an active hiking/singing/rationality/tech community?  
Does it attract many [group]?

Mastery

What opportunities would I have to learn new subjects outside of software?  
Who would I be learning from? (eg is there any mentoring/training?)  
How masterful are the people I would be working with?  
How much time would I be actively devoting to learning? (eg via experiments)

Autonomy

How much choice would I have in what to work on? (eg do I need permission)  
How much power would I have to make important decisions?  
How much would I be judged on results over process?  
(eg can I go home early if I'm not productive? can I take a walk instead?)  
How often would I be able to take time off to travel?  
How much free time would I have for personal projects?  
How flexible are the hours?

Purpose

What is the goal of this work?  
What are the odds of success?  
How is progress towards the goal measured? (ie how do you know if working)  
What is the process for making decisions? (eg is it based on evidence? intuition)  
Are there pressures towards / risks of instrumentally irrational decisions?

checklists

Fermi equations,  
cost/benefit analysis,  
priority list

improve cognition by preventing mistakes that happen without conscious awareness, what

Principles

Scepticism

One, what are they saying? Two, are they telling the truth? Three, does their reasoning make sense?

Honesty

Prioritise

Decouple

Decoupling vs Contextualising

Decoupling: looking at a single issue/question/idea/fact at a time.

Related ideas, implications and associations only brought in explicitly and

Contextualizing, on the other hand, means that all associative connections between

Be intentional

Maintain an exciting and compelling vision of the future to fight apathy.

Make sure it's slightly out of reach to keep pushing you. Never give up.

learn tools for a job, not for the tools

technical solutions to human problems

of which the classic example is that if you don't like your partner making noise, ignore them.

Noticing confusion

ALWAYS ASK FOR A COMPARATOR

reference class / baseline / dummy

Community

EA - Massive flow of great ideas, positivity, support, correction

Scientists and AI

Aligners

Exercise

mind needs body to move

3 x 3 x 3

Three priorities for the quarter, the week and the day.

Inventories

Oxford Utilitarianism

Morality as Cooperation

Big Five

Schwartz Values

Seligman "VIA"

Regular solitude

Comparing down / Gratitude

hap data

Stoicism

Get over yourself

Do not resent CHAI, do not shy

Avoid competitive suffering

Rest Day

Keeping slightly too busy

No time for rumination

Enables structured procrastination

Always have two things

Most often studying at work

One top idea

<http://www.paulgraham.com/top.html>

Acronym mnemonics

Murphyjitsu

## Environment

### Biology

- Short nails
  - anti-biting
- Glasses (;
- Sleep hygiene
  - Air quality
    - window open
    - Succulents?
  - Oxymetazoline for nose
  - Screen off at 11pm
  - Eyemask
  - Earplugs
- Biking + weights + yoga
  - Cardio + resistance + flexibility/mindfulness
  - + swimming for joints?
  - + cello for dexterity
- Complete meals
  - Time (-15mins prep? -5mins shop?)
  - Money (-£3 each?)
    - Queal costs: £252 for 1 light year
      - April 2018: £0
      - June: £48
      - September: £44
      - October: £45
      - Feb 2019: £55
      - Mar 2019: £60
- \* 360
- Delivery system for other powders
- optimal biochemistry
  - 0.25g citicoline
  - 0.2g theanine
  - 0.5 g ash
  - 1g green tea
- Chemical goals
  - Cognition
    - Nicotine
    - Choline
  - Sleep
    - Melatonin
  - Inflammation
    - Aspirin
    - Omega-3
  - Curcumin - black pepper
- Neuroprotection

Green tea  
Longevity (oxidation?)  
Vitamin K  
Green tea  
Vitamin D  
Vitamin B12  
Carnitine?  
Genetic diagnoses

#### Compounding

knowledge and skills,  
saving and investing,  
resistance training,  
networking

#### Triggers (Habits)

Up - pills, 10mins fun laptop  
Out door - pockets, book, queal  
Out door - bike.  
Remember Andreas' mockery  
After standup - deep work  
Lunch - phone K  
Home - empty pockets next to door.  
No fun laptop til 10pm.  
10pm - Cortisol Manager. Mix queal  
Bike in front of door  
Bed - no laptop, lights off  
Teeth  
Saturday morning - Weights, file nails

#### the lives of others

seek out magicians  
, and crack their magic  
imagine yourself as one, older and wiser, in great detail. Imagine  
yourself as the person you would be afraid to say you want to be out  
loud to others (because it seems so ridiculously impossible right now).  
Write it down in great clarity and detail, then forget it. And let the  
part of your subconscious mind that still remembers lead you to becoming  
the things you want, and maybe, years later, check if it did.

Greek chorus

Love

#### Finances

Saving  
Autosavings  
£6000 p.a.

Queal  
£3000 p.a. ?  
(£7 -> £1 each time)

Bike  
£7 each time  
£1400 p.a.  
More like £700 in practice

No booze  
£1500 p.a.

Own stacks  
£200 p.a.

Libgen  
£1000 p.a. ?

Torrents  
£500 p.a.

Bulk buys  
+ Rent...  
£6000 p.a.  
+ £2000 bills?  
Housemates  
+ Markets...  
Buy the most expensive gift in a cheap category  
£20 candle

+ Mustache  
Become a producer rather than a consumer  
Sustenance comes from your backpack  
Live a local life  
Don't commute  
Windfalls are for buying freedom  
You never, ever go "shopping"  
Stock up on sales  
You don't need much

Index funds  
<http://www.bayesianinvestor.com/blog/index.php/2015/07/28/advice-for-buy-and-hold-investing>

Barbell

Quantified self  
"Hap"  
RescueTime  
+ Oura / SleepScore / Fitbit Alta HR / GO2SLEEP  
ALEXFERGUS \$50  
+ Thriva  
+ Scales

Anti-Skinneboxing

Trying to work with an internet device is like dieting with beer & ice cream on your desk  
Against shallow content, against negative media  
No smartphone  
etc/hosts block  
ImpulseBlocker  
RescueTime  
BIMODAL: Lockdown laptop + Fuckaround device

Public book reviews  
leverage my vanity and Goodreads gamification to read more, read better

#### Info

Rats Twitter  
Rats Goodreads  
Gelman  
Libgen ()  
SciHub  
More or less

#### Device

PDFs, 2FA, Maps, Taxi GPS  
Redundant convergence: Books, music/Podcasts, camera, mic, emergency payment NFC, budget

#### Two big screens

Antilibrary for tangible humility

#### Infosec

gleech.org/browser

#### Weirdness points

{% include process/critik.html %}  
{% include comments.html %}

# Revolutionary progress outside universities

Gavin

2019-07-17

Two large facts:

Lots of people are very jaded about academic research.

The entire history of science up to about 1750 and large parts until 1950 Post Office Research Station and Bletchley, OP-20-G (computer science, crypto and crypta) Manhattan Project (nuclear physics) Telecommunications Research Establishment (operations research) Santa Fe (complexity) Bell Labs (information theory) [https://en.wikipedia.org/wiki/Janelia\\_Research\\_Campus](https://en.wikipedia.org/wiki/Janelia_Research_Campus) <https://patrickcollison.com/labs> RAND (game theory) ARPA (networks) ONI (cybersecurity) <https://ea.greaterwrong.com/posts/dCjz5mgQdiv57wWGz/ingredients-for-creating-disruptive-research-teams> Has anyone done what MIRI is trying to? Thinktanks Fabians neolibs Neocons [https://en.wikipedia.org/wiki/International\\_Institute\\_for\\_Applied\\_Systems\\_Analysis](https://en.wikipedia.org/wiki/International_Institute_for_Applied_Systems_Analysis) [https://www.aeaweb.org/rfe/showCat.php?cat\\_id=44](https://www.aeaweb.org/rfe/showCat.php?cat_id=44)

# The most useful books

Gavin

2019-08-09

A typical American film, naive and silly, can — for all its silliness and even by means of it — be instructive. A fatuous, self-conscious English film can teach one nothing. I have often learned a lesson from a silly American film.

— Wittgenstein

There's nothing so practical as good theory.

— Kurt Lewin

*Attention conservation notice:* All of these can be summarised, all of these can be reinvented, all of these are unnecessary for some people. (e.g. the naturally happy or productive).

## ***Scarcity***

## ***Practical biochemistry***

I've not found a good book about supplementation. The paid parts of Exam-  
ine.com are the closest thing.

Why care? Most people away from the equator vit D Dave Gwern Ssc dep anx  
Melatonin

## ***Exercise***

- Starting Strength

## ***Probability***

I mostly left out maths and technical matters, but some understanding of probability is essential for anyone.

Yud Fooled by Randomness Algorithms to Live By How Not to Be wrong

*Little Book of Common Sense Investing*

*Cybersecurity*

Schneier  
Mitnick

*Scripting*

Automate the Boring stuff  
Regex  
Scraping  
Scripting

*Modelling*

Doing Data Science

*Work*

80k  
Never Split the Difference  
Antianxiety  
Elephant in the Brain  
Peace of understanding, amor dei int  
Or Waking Up  
Taleb / Kahne / Rosling / Galef  
Against news  
How to Talk About Books

*Anki Essentials*

*Productivity*

Perry GTD Eat that Frog

# What is best?

Gavin

2018-08-23

It is insufficient to protect ourselves with laws; we need to protect ourselves with mathematics.

- Bruce Schneier

You are objectively correct to be very uncertain about the particular risk to attack. My decision to focus on AI among risks is a “strong opinion weakly held”, to facilitate moving forward in roughly the right direction.

The most-modest prior (just taking the median from the surveys of mainstream AI experts) on  
The weak evidence of expert surveys: 4% risk of major AI disaster this century  
It's still more neglected than bio, nuclear, or speculative climate things, etc.  
And the bottleneck is full-time senior researchers, not e2g money or essayists  
Much better optionality than these others (if I prove to dislike research, or if some other  
And it also gives internal options: It's pretty common to quit a PhD midway and suffer m  
Doing a technical AI PhD lets you move into policy easily  
AI is the only really plausible source for s-risk, which I take quite seriously.  
Can monitor capabilities labs while making giant pots of cash  
Biorisk is really hard to help with if you're not in government or wet labs  
I find it more interesting and closer to my skillset  
I want to help shift the field away from complacency  
I love research and do it compulsively  
I went for a couple of ML engineer posts (CHAI, Ought) and didn't get them.

I suppose a more robust pick would be cause prioritisation research or FHI style macrostrategy. But that's very difficult to get into - unless you happen to have a suffered-focussed ethics (which is glutted with money), which I don't. And the optionality is low.

How did you decide to focus on AI? I took about 3 years to accept the cause (after reading Superintelligence) because I thought MIRI was the only game in town, and their focus on proofs and no-feedback design (which they've since relented on) struck me as doomed. But actually there are several very different methodologies (decision theory, ML/RL, question-answering) and a dozen agendas for AGI Safety, including a mainstream iterable RL one at DeepMind, OpenAI, and Ought.

How did you decide to work on AI directly over earning-to-give with programmer money? Spirit of exploration plus excellent options if I stop.

How did you decide to start a PhD instead of doing a “software engineer at an AI organization” role? Tried, but didn’t get it. I’ll study more whiteboard algos next time. (; I was actually agnostic about Eng vs PhD, but trying engineer work seemed less of an investment / more fail-fast than the PhD.

I should have said “main bottleneck”; it sure seems like there is a shortage of aligned ML engineers too (CHAI were looking for one for more than a year). That line just speaks against prioritising e2g for AIS. Z&O’s perspective looks good to me - was there something in particular?

Opinions really differ on PhDs. I know some people (MIRI - CFARish) who think that the incentives are bad enough that you’re better off getting grants and teaching yourself. But e.g. DeepMind and OpenAI value them fairly (but not excessively) highly. And but we will need academics to - at very least! - make legible the breakthroughs of the mad monks.

I have a Plan, but it is perhaps uselessly disjunctive: “I’ll get my PhD [or not], then work on safety at a low-replaceability place, possibly the Turing Institute or DeepMind, and so try to ensure that there is at least one person-who-knows-what-backprop-is in the room where the policy decisions are made. In Britain. [or I’ll drop out and work at a normal FANG lab, making dozens of small grants to junior researchers a year and keeping an eye on them.]”

Have you asked for info on EA Corner Discord? Some distinguished people on there, answering all kinds of questions. I flatter myself that by taking this funded place I’m preventing ‘one unit’ of thoughtless capabilities research. I don’t know the value of that, for the average low-impact researcher. I am haunted by the idea that researcher impact is heavy-tailed, and that I am unlikely to be in that tail. But I’ll roll and see.

“So it sounds like you decided on working on AI for all the reasons mentioned above, and then in the world of AI you saw two tracks: research engineer (no PhD required) or researcher (PhD required). And you were interested in the research engineer positions, but those didn’t work out, so you decided to go down the PhD route. Is that right?”

Some questions: - You say “the bottleneck is full-time senior researchers”; does this mean that, by contrast, you think a research engineer position has pretty high replaceability? - Do you have a Nate Soares-style plan? (i.e. “I’m going to get my PhD, then work on safety at OpenAI, and try to ensure that OpenAI is the first organization in the world to build powerful AGI.”) Do you recommend it? - Have you listened to the 80k podcast with Daniel Ziegler and Catherine Olsson? What do you think of it and of Catherine’s advice about PhDs in AI safety?

# Which AI safety agenda?

Gavin

2019-08-23

{% assign src = "https://www.lesswrong.com/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment" %} {% assign = "" %}

*The following is based entirely off Shah and Cottier's excellent work. Go read it first.*

The idea is closely connected to the problem of artificial systems optimizing adversarially against humans. The idea must be explained sufficiently well that we believe it is plausible.

1. orthogonality
2. complexity of value
3. Goodhart's Curse
4. Will AI be deployed in places it can cause catastrophe?
5. Agentive AGI?

Will the first AGI be most effectively modelled like a unitary, unbounded, goal-directed agent?

2. Incentive for agentive AGI?

Would unitary goal-directed agents have a worthwhile advantage over other superintelligent systems?

3. Modularity over integration?

In general and holding resources constant, is a collection of modular AI systems with distinct interfaces more competent than a single integrated AI system?

**Related reading:** Reframing Superintelligence Ch. 12, 13, AGI will drastically increase economic power  
**Comment:** an almost equivalent trade-off here is generality vs. specialisation. Modular systems

4. Does current AI R&D extrapolate to AI services?

AI systems so far generally lack some key qualities that are traditionally supposed of AGI, namely: pursuing cross-domain long-term goals, having broad capabilities, and being persistent and unitary. Does this lacking extrapolate, with increasing

automation of AI R&D and the rise of a broad collection of superintelligent services?

5. Incidental agentive AGI?

Will systems built like unitary goal-directed agents develop incidentally from something humans or other AI systems build?

6. Convergent rationality?

Given sufficient capacity, does an AI system converge on rational agency and consequentialism to achieve its objective?

7. Mesa-optimisation?

Will there be optimisation processes that, in turn, develop considerably powerful optimisers to achieve their objective? A historical example is natural selection optimising for reproductive fitness to make humans. Humans may have good reproductive fitness, but optimise for other things such as pleasure even when this diverges from fitness.

8. Recursive self improvement?

Is an AI system that improves through its own AI R&D and self-modification capabilities more likely than distributed AI R&D automation? Recursive improvement would give some form of explosive growth, and so could result in unprecedented gains in intelligence.

9. Discontinuity to AGI?

Will there be discontinuous, explosive growth in AI capabilities to reach the first agentive AGI? A discontinuity reduces the opportunity to correct course. Before AGI it seems most likely to result from a qualitative change in learning curve, due to an algorithmic insight, architectural change or scale-up in resource utilisation.

10. Discontinuity from AGI?

Will there be discontinuous, explosive growth in AI capabilities after agentive AGI? A discontinuity reduces the opportunity to correct course. After AGI it seems most likely to result from a recursive improvement capability.

11. ML scales to AGI?

Do contemporary machine learning techniques scale to general human level (and beyond)? The state-of-the-art experimental research aiming towards AGI is characterised by a set of theoretical assumptions, such as reinforcement learning and probabilistic inference. Does this paradigm readily scale to general human-level capabilities without fundamental changes in the assumptions or methods?

12. Deep insights needed?

Do we need a much deeper understanding of intelligence to build an aligned AI?

### 13. Broad basin for corrigibility?

Do corrigible AI systems have a broad basin of attraction to intent alignment? Corrigible AI tries to help an overseer. It acts to improve its model of the overseer's preferences, and is incentivised to make sure any subsystems it creates are aligned — perhaps even more so than itself. In this way, perturbations or errors in alignment tend to be corrected, and it takes a large perturbation to move out of this “basin” of corrigibility.

### 14. Inconspicuous failure?

Will a concrete, catastrophic AI failure be overwhelmingly hard to recognise or anticipate? For certain kinds of advanced AI systems (namely the goal-directed type), it seems that short of near proof-level assurances, all safeguards are thwarted by the nearest unblocked strategy. Such AI may also be incentivised for deception and manipulation towards a treacherous turn. Or, in a machine learning framing, it would be very difficult to make such AI robust to distributional shift. Related reading: Importance of new mathematical foundations to avoid inconspicuous failure

Agentive -> -> Proof-level assurance

### 15. Creeping failure?

Would gradual gains in the influence of AI allow small problems to accumulate to catastrophe? The gradual aspect affords opportunity to recognise failures and think about solutions. Yet for any given incremental change in the use of AI, the economic incentives could outweigh the problems, such that we become more entangled in, and reliant on, a complex system that can collapse suddenly or drift from our values.

# 'The Odyssey' (2017) by Emily Wilson

Gavin

2019-09-03

{% assign egan = "https://en.wikipedia.org/wiki/The\_Moral\_Virologist" %}

I don't want to hector Homer, but somehow this was both boring and evil, childish and didactic. I won't belabour the book's immorality, since it's so obvious; it's the near-total absence of artistic merit that is not obvious. I found nothing in it worth reading or quoting until Book 9, nearly halfway through. These are songs of praise of warmongering pirates. (People love pirates, and I say let em. Just don't call them paragons.)

The ideology is dad porn, a set of thin, obvious, animal values. "Kings do whatever they want - death for messing with a noble; don't cross the priests; offer huge sacrifices; always do what your husband and dad say; the unlucky and the disabled are cursed and to be shunned; blood is blood is blood." (It's not as if they could easily have been otherwise. Too poor, too lawless and misruled, too near to nature.)

The ghost of Agamemnon answered, "Lucky you, cunning Odysseus: you got yourself a wife of virtue—great Penelope. How principled she was, that she remembered her husband all those years! Her fame will live forever, and the deathless gods will make a poem to delight all those on earth about intelligent Penelope.

(Odysseus sleeps with half a dozen other women and demigods, most of them begging him to, and needless to say suffers nothing of it.)

There's no mention of the suffering of the several cities he sacks, or the many tacitly raped women. Dozens of people are murdered for being rude, though. For a quasi-sacred text there's a surprising amount of unpunished priest killing (e.g. Leodes).

The structure is awful: we see almost nothing of Odysseus for the first quarter of the poem, instead following his son around as he listens to a series of boring old men. Most of Odysseus' feats are not shown, are instead related by him as unaffectionate stories. (I suppose we could amuse ourselves by treating this as unreliable narration, but they certainly didn't.) And the poem doesn't end at its climax, instead meandering on through another few books of pointless back-patting.

(Should I go easy? After all, this is groundbreaking work, the prototype of art. Sure; I'll go easy if you stop hyping it and making everyone read it as an exemplar.)

It must be a cliche among classicists that the ‘Classical’ civilisations were not classical in the sense of being austere, logical, tasteful, or contemplative. That they were not Apollonian, that only a handful of people in them were. I hope my rant here is not just me being misled by the modern sense of “hero” - but the fact is that Odysseus wins, is praised endlessly, and his rights trump all else.

This isn’t just me being clueless, post-oral, and close-minded: The ancients were well aware that the ending is unsatisfying crap. One popular headcanon was that, after Odysseus slays the suitors, he is immediately exiled from Ithaca, set adrift again. Cue the music!

---

One reading of Odysseus’ name is as variant of the verb ‘to be hated’. So a calque might be “King Punchable of Ithaca”. (“the most unhappy man alive”)

Odysseus is treated incredibly well by almost everyone, despite his crimes. Complete strangers oil him up and dress him in fine “woolen cloak and tunic” eleven times, and he is given precious weaponry and potions for nothing several times. This is supposed to reflect on him, but instead it shows the Greek ideal of hospitality, one of the few nice things in that culture.

He appears to sincerely miss Ithaca (his status more than his wife), weeping frequently. But he also fucks about all the time, for instance staying an entire year voluntarily enjoying Circe.

It is completely unclear what O does to deserve his fortune. (Whereas his misfortune is always directly linked to his own machismo or idiocy.) The only virtues we see him exercise directly (not counting brute aggression and discus throwing) are courage and cunning (specifically lying). Ok, he also makes one good speech:

‘Listen to me, my friends, despite your grief. We do not know where darkness lives, nor dawn, nor where the sun that shines upon the world goes underneath the earth, nor where it rises. We need a way to fix our current plight, but I do not know how…’

I suppose we can put the rest down to charisma, the oddest and least rational of human powers.

‘It seems that everybody loves this man, and honors him, in every place we sail to.’

Everyone extols him without him ever demonstrating the virtues they extol. (Politeness, propriety, wisdom, strategy...) Every other idiot is “godlike” at something or other, and seeing the state of their gods you see how this could be true. At least it’s funny:

He went out of his bedroom like a god  
King Menelaus, you are right... Your voice is like a god's to us.  
Majestic, holy King Alcinous leapt out of bed, as did Odysseus the city-sacker.  
Then the blessed king, mighty Alcinous, led out his guest...

(The gods are stupid mirrors of Greek nobility; for instance they have supernatural slaves, the nymphs.) This at least is a philosophical difference between them and I: in their superstitious idealist mode, properties aren't for describing the present, but instead the timeless essence of a thing. Wilson:

Ships are "black", "hollow", "swift" or "curved", never "brown", "slow" or "wobbly"... Penelope is "prudent Penelope", never "swift-footed Penelope", even if she is moving quickly. Telemachus is thoughtful, even when he seems particularly immature.

All the feats of the heroes are totally dependent on the power of gods. If they say you can't sail, you can't.

His skin would have been ripped away, and his bones smashed had not Athena given him a thought.

Athena poured unearthly charm upon his head and shoulders, and she made him taller and sturdier, so these Phaeceans would welcome and respect him.

Without Hermes or Athena constantly intervening, O would be nowhere, achieve nothing. One nice tension here though:

But death is universal. Even gods cannot protect the people that they love, when fate and cruel death catch up with them.

One of the few times I felt sympathy for Odysseus was when he was trying to lead his men, who are mainly large-adult-sons. (Same with the suitors.) One breaks his neck falling down a ladder. They undo a month of work by playing with the bag of winds. Several times they are totally paralysed by their wailing and tantrums.

As when a herd of cows is coming back from pasture into the yard; and all the little heifers jump from their pens to skip and run towards their mothers, and they cluster round them, mooing; just so my men, as soon they saw me, began to weep...

The other men... wept for those that died. I ordered them to stop their crying, scowling hard at each.

Odysseus occasionally draws his sword on them for backtalking him, or running around like Muppets. Their deaths are roughly equally due to Odysseus' aggression and avarice, and their own foolishness.

I cheered the uprising against him, who are completely in the right. But of course they lose, because of mere divine intervention.

---

OK I lied: I will talk about evil. Though by the end of this I was jaded and dismissive, the aftermath of Odysseus slaughtering the suitors still struck me as an atrocity unusual for the genre:

“When the whole house is set in proper order, restore my halls to health: take out the [slave] girls between the courtyard wall and the rotunda. Hack at them with long swords, eradicate all life from them. They will forget the things the suitors made them do with them in secret, through Aphrodite...” “I refuse to grant these girls a clean death, since they poured down shame on me and Mother, when they lay beside the suitors.” At that, he would a piece of sailor’s rope round the rotunda... just so the girls, their heads all in a row, were strung up with the noose around their necks to make their death an agony. They gasped, feet twitching for a while, but not for long.

I’ve read de Sade, Kaczynski, Himmler, Houellebecq, Egan and Watts at their most dyspeptic; it’s not that I’m squeamish about real or fictional evil, or that my sulking sense of justice blinds me to aesthetics. This sort of thing happened; nothing cannot be said; maybe even nothing cannot be said beautifully. It’s just that, again, there is nearly no nobility and no classicism in this. I am so glad this culture is gone.

---

Did its audience know the story was bullshit? Or was it scripture to them? (Like most scripture, it is pathetically ignoble, violent, and self-serving.) Well, they don’t seem to have had scripture, not even Hesiod. So Homer is more like Dante or Milton for them: not sacred, but pious and moralising.

How big was mighty Troy? How noble was godlike Odysseus? How petty their pantheon? How long this epic?

- 
- Even thought-provoking bits like the lotus eaters or Cyclopean anarchism are over in less than half a page.
  - Surprised when Zeus was described as “husband of Hera”.
  - The “no man” pun thing was so stupid I had to put the book down for a couple of days.

---

Normally I would stop reading a book this bad, but I read it to prepare for Ulysses, so I dragged myself through.

I don’t think the badness is due to Wilson. I actually quite like her style, and it’s the skeleton of plot, sentiment, and moral that repulses me.

Her introduction takes up a quarter of the entire book. It’s good and sane but repetitive, taking pains to spell out all the ignoble and questionable, all the ugly

and clumsy parts. I don't know how she keeps up her enthusiasm for the book, in the face of them, but more power to her.

<h3>One man's modus ponens is another man's modus tollens</h3>  
<div>

You can read the above as a demonstration of my lack of taste: if every prof on earth says

Maybe I just need to read another, less spartan translation. But then it would be Chapman's

<!-- -->

Maybe I'd get it if I read Bloom's book about it. But it's longer than the original work.

<!-- -->

I think it's mostly likely a missing mood of mine. I don't even vaguely sympathise with

</div>

# Notes on OpenAI Five

Gavin

2019-08-23

OpenAI Five has hardcoded Dota knowledge in its reward. Two readings of “hardcoded”:

1. initialised by a human;
2. fixed by a human (i.e. no updates to R from self-play).

Both readings obtain here.

A strict definition of “hardcoded reward knowledge”: if the reward function includes human decisions about anything but {positive reward for winning} and {negative reward for losing}, it has hardcoded reward knowledge.

(It’s plausible that less strict definitions are fairer, e.g. in this case the software agents are handicapped by not using intra-team communication, so reward shaping to simulate communication - e.g. lane assignment - could be seen as *fair* hardcoding.)

*Domain-specific manual reward-feature selection*: The game API reports 20,000 features. The handcrafted reward function includes only 28 (17 + 7 building healths + lanes). On top of the feature selection, the weights of each of these features are also handcrafted!

Take “reward shaping” to mean supplementing or replacing the natural endpoint rewards (team win and team loss) with domain-specific intermediate rewards selected by a human. OAI5’s reward is completely “designed by [OpenAI’s] local Dota [human] experts”, including selecting a tiny fraction of the most important features and setting the weights of the features, so it has domain-specific hardcoded knowledge.

The reward processing used is non-domain-specific, since it would apply to any mixed co-operative / competitive game.

---

That covers hardcoded knowledge in *the reward function*. Another vector for hardcoding is the inductive bias of the architecture used: we search a huge number of ANN structures to find a particular Dota-friendly one. I’m ambivalent about whether this counts as hardcoding, and ignore it in the following.

Another kind of hardcoded, but uselessly intractable would be manually tinkering with e.g. buggy activation functions, e.g. using model explanation to select individual nodes. It is vanishingly unlikely that OpenAI did this.

So my definition of hardcoded is "some degree of at least one of

- a subset of features are selected by humans
  - feature rewards are fixed by humans
  - post-hoc manual edits are made to the network."
-

# Turing vs Plato

Gavin

2019-11-23

in all sciences, as in plain mirrors, some marks and images of the truth of intelligible objects appear, but in geometry chiefly; which... doth bring back and turn the understanding, as it were, purged and gently loosened from [the mere] senses... Plato himself dislikes Eudoxus, Archytas, and Menaechmus for endeavoring to bring down the doubling of the cube to mechanical operations; for by this means all that was good in geometry would be lost and corrupted; it falling back to sensible things, and not rising upward and considering immaterial and immortal images, in which God being versed is always God.

– Plutarch

[Tarski and I both stress] the great importance of the concept of... Turing's computability... this importance is largely due to the fact that, with this concept, one has for the first time succeeded in giving an absolute notion to an interesting epistemological notion, i.e., one not depending on the formalism chosen.

– Gödel

equivalence of mental objects

physical instantiation of mental objects

“Turing is a nightmare for Platonists. The point of an ideal is that it's separate from the shadows on the wall - that its nature is contained within itself. Turing shows that in a universal machine, any form of reasoning can emerge through determinate rules...”

We know now that there are infinitely many such machines. Thus, for any ideal you present to me, I can find an equivalent representation of it using different terms. This means that the uniqueness of an ideal is not given by its singular form, but by the constraints imposed on it by other ideals - i.e. invariants/theorems which hold regardless of the representation chosen.

In this light, the platonist has three unattractive options:

1. formulate a new Cave metaphor which captures the suspicion about the phenomenal world's unreliability while granting the underdetermination of reason: that there is no one true shape, that it's shadows all the way down.
2. accept that logical distinctness is illusory or at least secondary, since everything is interlinked (by computability). This is really just Pre-Socratic mysticism again.
3. realism. The shadows are all there is. Ideals are only tools.

I actually still haven't joined the dots here. There is either something wrong, or just missing from the argument.

Placeholders:

<https://schwitzsplinters.blogspot.co.uk/2017/09/how-to-build-immaterial-computer.html> <https://arxiv.org/pdf/math/0209332v1.pdf> <https://philpapers.org/rec/RANCP>

---

### Soul as the Incomputable

The refuge of scoundrels is the incomputable. Penrose Goedel brains. Mathematical creativity.

So we could reply:

*This post is just context for ideas by John Morrice.*

# The trouble with supplements

Gavin

2019-12-10

{% include pills/links.html %}

there may be useful interventions, but they will be of little value on average — if the benefit is universal, then it will be small; if it is large and predictable, then it will be limited to the few with a particular disease; otherwise, it will be unpredictably idiosyncratic so those who need it will not know it. Thus, the metallic laws: the larger the change you expect, the less likely it is; the low-hanging fruit, having already been plucked, will not be tested; and the more rigorously you test the leftovers, the smaller the final net effects will be.

• Gwern Branwen

every mouthful of food you and I have ever taken contained many billions of kinds of complex molecules whose structure and physiological effects have never been determined – and many millions of which would be toxic or fatal in large doses... we are daily ingesting thousands of substances that are far more dangerous than saccharin – but in amounts that are safe, because they are far below the various thresholds of toxicity. At present, there are hardly any substances, except some common drugs, for which we actually know the threshold.

• Edwin Jaynes (...)

In the last century, half a revolution happened: you can now buy many thousands of substances that claim to promote health, and perhaps a couple of them do.

The point is to fine-tune your health: to prevent idiosyncratic disorders, to treat ubiquitous “subclinical” or “subsyndromal” or (worst of all) “idiopathic” health problems. All the little things that make life worse. And so most UK adults take supplements (about half of those multivitamins). Several problems with this:

1. *Absence of general evidence / Evidence of harm.* Many supposedly health-promoting substances have uselessly weak evidence. For instance, frequent use of multivitamins is probably somewhat harmful: they *increase* mortality for the average user, maybe due to overdosing you with antioxidants.

2. *Physiology is personal.* Even for substances that have general warrant, the ‘heterogeneity’ in their effects and side-effects can be enormous, even for quite closely matched pairs. (For instance, some people don’t get *any* stimulation from caffeine for genetic reasons. Morphine, the central example of a powerful and basic drug, has a “number needed to treat” post-op pain of 2.9 - i.e. on average a high dose only works well for one in three people!)
3. *Geographical and seasonal variation:* for instance, during winter, around a third of UK adults are deficient in vitamin D.
4. *Snake oil on the margin:* The supplement industry is regulated (for instance, you have to apply to the European Commission if you want to make a health claim for your product), but misleading claims and inaccurate concentrations are common. (For instance, the Ayurvedic supplement bacopa has been known to contain unsafe levels of lead and other heavy metals.)
5. Powdered supplements are often 2-4x cheaper than pills, but are fiddly and sometimes taste bad. 1
6. *What counts is latent:* There are now cheap places to get blood tests (or genome hits) for particular biomarkers, which you’d think would close the gap. But blood markers are only proxies for the *real* target variables: mortality, productivity, mood, etc.

The missing half of the revolution is measurement. The sensible supplementer needs three kinds of data to avoid waste and unnecessary risk:

- general clinical findings,
- personal experiments (biochemistry before and after, control doses, measurements of actually valuable responses),
- chemical assays of particular products.

A shame that general clinical findings are so unreliable, and that getting strong personal data remains the province of heroically nerdy people, willing to invest dozens of hours into self-experimentation (reading papers, double-blinding with self-filled capsules, data collection), including learning how to analyse the results sensibly. There’s something sad about this: that external validity is so hard in biomed (and society...) that all we can really trust is local inference, n-of-1 description.

Despite plummeting measurement costs (blood tests down by 100x, genome sequencing down by 100,000x, all the consumer QS gizmos), the money and time required for an actually-scientific supplement habit are still prohibitive. So: you take safe inexpensive things and live with the uncertainty - or, more, you rely on a prior that evolution is hard to beat on body matters, and lean against taking anything except the most convincing substances.

There are economies of scale to summarising and operationalising research, testing

batches, and filling capsules. And removing gatekeeping for cheap important tests has the benefit of raising our autonomy, putting us in control of at least the minor things. So is this a gap in the market? I don't really know, I just want it to exist. (There are already well-funded toy versions of a personalised service, but their offering is pretty superficial so far.)

```
<h3>Another general counterargument</h3>
<div>
    There is sometimes value in mere sufficiency. Across species, across phyla, there seems
    <blockquote><i>
        When there is an abundance of nutrients, the signal is to focus on reproduction,
    </i></blockquote><br>
    The underlying claim is something like "metabolism is violent, so things which boost it
    So aiming to close all gaps - calories, amino acids, antioxidants - may end up having <%
</div>
{%
    include pills/foots.html
%}
```

# Nation playlists

Gavin

2020-09-02

```
{% include nation/links.md %}

<h3>uk</h3>
<div>
    <div class="accordion">
        <h3>Scotland</h3>
        <div>
            {% include nation/scot.html %}
        </div>
        <!-- -->
        <h3>England</h3>
        <div>
            {% include nation/eng.html %}
        </div>
        <!-- -->
        <h3>Wales</h3>
        <div>
            {% include nation/wales.html %}
        </div>
        <!-- -->
        <h3>Northern Ireland</h3>
        <div>
            {% include nation/ulster.html %}
        </div>
        <!-- -->
        </div>
    </div>
    <!-- -->
    <!-- -->
    <h3>americas</h3>
    <div>
        <div class="accordion">
            <h3>Brazil</h3>
            <div>
                {% include nation/bra.html %}
            </div>
        </div>
    </div>
```

```
</div>
<!-- -->
<h3>Canada</h3>
<div>
    {%
        include nation/can.html %
    }
</div>
<!-- -->
<h3>America</h3>
<div>
    {%
        include nation/us.html %
    }
</div>
<!-- -->
<!-- -->
<h3>euro</h3>
<div>
    <div class="accordion">
        <!-- -->
        <h3>France</h3>
        <div>
            {%
                include nation/fra.html %
            }
        </div>
        <!-- -->
        <h3>Germany</h3>
        <div>
            {%
                include nation/ger.html %
            }
        </div>
        <!-- -->
        <h3>Italy</h3>
        <div>
            {%
                include nation/ita.html %
            }
        </div>
        <!-- -->
        <h3>Netherlands</h3>
        <div>
            {%
                include nation/neder.html %
            }
        </div>
        <!-- -->
        <h3>Estonia</h3>
        <div>
            {%
                include nation/est.html %
            }
        </div>
        <!-- -->
        <h3>Hungary</h3>
```

```
<div>
    {%
        include nation/hun.html %}
</div>
<!-- -->
<h3>Spain</h3>
<div>
    {%
        include nation/esp.html %}
</div>
<!-- -->
<h3>Portugal</h3>
<div>
    {%
        include nation/portug.html  %}
</div>
<!-- -->
</div>
</div>
<!-- -->
<!-- -->
<h3>asia</h3>
<div>
    <div class="accordion">
        <!-- -->
        <h3>Japan</h3>
        <div>
            {%
                include nation/jp.html  %}
        </div>
        <!-- -->
        <h3>Mainland China</h3>
        <div>
            {%
                include nation/chin.html      %}
        </div>
        <!-- -->
        <h3>Hong Kong</h3>
        <div>
            {%
                include nation/hk.html  %}
        </div>
        <!-- -->
        </div>
    </div>
    <!-- -->
    <!-- -->
    <h3>oceania</h3>
    <div>
        <div class="accordion">
            <!-- -->
            <h3>Australia</h3>
```

```
<div>
    {%
        include nation/aus.html %}
    </div>
</div>
```

## Caveats

Alright, so it is weird to do one list for all of Brazil (pop. 210m) and one for Bristol (pop 0.6m). I welcome contributions from scholars of the music of Feira de Santana (or indeed from any place on earth).

My selection from non-Anglophone countries will be biased towards obviousness and against wit. I forgive a gifted lyricist almost anything (for instance, I love early Mountain Goats), and I mostly can't here.

Most countries seem to have local Indie Gods: Tragically Hip (Canada), Microdisney (Ireland). Mostly don't survive leaving their context.

# Better ways to write maths

Gavin

2020-09-26

```
{% include phone_img.html %} {% assign adm = "https://en.wikipedia.org/wiki/De_Morgan%27s_laws"
%} {% assign jang = "https://blog.evjang.com/2018/08/dijkstras.html" %} {% assign sym = "https://en.wikipedia.org/wiki/History_of_mathematical_notation#Symbolic_stage"
%} {% assign tao = "https://mathoverflow.net/questions/366070/what-are-the-benefits-of-writing-vector-inner-products-as-langle-u-v-rangle/366118#366118"
%} {% assign tao2 = "https://terrytao.wordpress.com/advice-on-writing-papers/use-good-notation/" %} {% assign qc = "https://quantum.country/" %}
{% assign di = "https://en.wikipedia.org/wiki/Dependency_inversion_principle"
%} {% assign color = "https://onlinelibrary.wiley.com/doi/pdf/10.1111/cxo.12676"
%} {% assign cole = "https://books.google.co.uk/books?id=BGM_hYKAgksC"
%} {% assign word = "https://en.wikipedia.org/wiki/Proof_without_words" %}
{% assign romer = "https://paulromer.net/jupyter-mathematica-and-the-future-of-the-research-paper/" %} {% assign href = "https://ctan.org/pkg/hyperref?lang=en"
%} {% assign mackay = "http://www.inference.org.uk/mackay/itila/book.html"
%} {% assign chen = "https://web.evanchen.cc/napkin.html" %} {% assign bret = "http://worrydream.com/ScientificCommunicationAsSequentialArt/" %} {% assign dist = "https://distill.pub/2020/communicating-with-interactive-articles/"
%} {% assign dist2 = "https://distill.pub/2020/circuits/zoom-in/" %} {% assign sipser = "https://www.goodreads.com/book/show/400716.Introduction_to_the_Theory_of_Computation"
%} {% assign qiao = "https://mobile.twitter.com/QiaochuYuan/status/1306035720109404162"
%}
```

... the contradictory opposite of a copulative proposition is a disjunctive proposition composed of the contradictory opposites of its parts... the contradictory opposite of a disjunctive proposition is a copulative proposition composed of the contradictories of the parts of the disjunctive proposition.

— William of Ockham (1355), or:

$$\sim(P \wedge Q) \rightarrow (\sim P \vee \sim Q) \sim(P \vee Q) \rightarrow (\sim P \wedge \sim Q)$$

— Augustus De Morgan (1860)

Any impatient student of mathematics or science or engineering who is irked by having algebraic symbolism thrust upon him should try to get along without it

for a week.

— Eric Temple Bell

Mathematical notation is not finished. You can tell, because so much of it is new, and because so many smart people struggle with it as it is.

Still, a set of conventions have hardened in the last 100 years. Maths is as terse as possible; monochrome; unfriendly; operates at full generality; and gives bad, undescriptive names to its objects.

Now, aside from the distress it causes the beginner, terseness is *good*: it lets us fit more in our head at once, and so go faster, and so go further. The move from prose to symbols is objectively an improvement, even as the appearance of maths moved further from human intuition.

What else is good about the conventional style? It is minimalist; it does not patronise; it is tasteful and grown-up; its generality saves a lot of ink; its leaving almost everything unsaid saves a lot of time. To master a conventional serious proof is to overcome an adversary, to simultaneously prove something about oneself.

Here are some different ways of doing it, less optimised for past masters.

## Colour

Use colours to instantly relate symbols to explanations, whether verbal or graphical. Like Eric Jang's incredible 'Dijkstras in Disguise':

This is also an instance of giving people several angles of attack on the same concept.

(There's mixed evidence about coloured text and comprehension in general, but the studies all focus on ordinary prose and I doubt they transfer to understanding formulae with dozens of symbols.)

## Comments

For example, you may come across definitions like this: "A finite state automaton is a quintuple (

$Q$

,

$\Sigma$

,

$q_0$

,

$F$

,

 $\delta$   
 $)$  where  $Q$  is a finite set of states  
 $q_0$   
 $,$   
 $q_1$   
 $, \dots,$   
 $q_n$   
 $),$   
 $\Sigma$   
is a finite alphabet of input symbols,

$q_0$   
is the start state,  
 $F$   
is the set of final states  
 $F \in Q$   
, and

$\delta \in Q \times \Sigma \times Q$   
, the transition function.”

That definition should be taken outside and shot.

~ John Coleman

rigour follows insight, and not vice versa.

~ James Stone

Michael Sipser has good comments on all the proofs in his great CS book:

```
<h3>Diagonalisation</h3>
<div>
  <center>
    
  </center>
</div>
```

Evan Chen’s book for bright highschoolers is suitably friendly too.

For learning material (rather than research writeups), the steps of a proof could be tagged as “routine”, “creative”, “tricky”, or “key” (h/t Qiaochu). These would be best as sidenotes.

Further: Why is there no metadata? The field dependencies; the theorem dependencies, upfront; how important this result is, for what; some proofs with a similar flavour; or, for fun, what’s the newest result necessary for this proof? When could it first have been proved?

## Motivating examples

A good stock of examples, as large as possible, is indispensable for a thorough understanding of any concept, and when I want to learn something new, I make it my first job to build one.

— Paul Halmos

Most maths writing jumps straight to the general definitions. But at least some people need to work up from examples and counterexamples instead.

This is another place that Chen's basic book beats high-status university texts: Literal examples are just one answer to the question "*Why should I care about this theory?*". Maybe authors think that question is wishy-washy, but examples are not subjective, just partial. I'm not even asking for — *horror of horrors!* — applications. Maybe generality feels strong: to solve all examples at once, without looking at them, is to rise above the objects.

There is an ignorant way of asking "Why should I care?": the way with no sense of aesthetics, curiosity, patience, the philistine way that cannot see any value without an application behind it, or money. This is maybe the way mathematicians take the question, and so maybe why they shun it.

## Composing subproofs

Here's proof by induction as an algorithm:

You then see that for any given instance you just need to write the two subroutines `BaseCase` and `InductiveStep`. I find this much easier to understand.

More generally I don't see much dependency inversion in proofs. Long proofs will include a sketch of the strategy, but mostly not with this lucidity. (Exceptions: Sipser, Chen.)

Maybe this only works if you know some programming before you do higher maths (a lamentably rare condition).

Here's an unfair but illuminating rant:

Imagine I asked you to learn a programming language where: - All the variable names were a single letter, and where programmers enjoyed using foreign alphabets, glyph variation and fonts to disambiguate their code from meaningless gibberish. - None of the functions were documented, and instead the API docs consisted of circular references to other pieces of similar code, often with the same names overloaded into multiple meanings, often impossible to Google. - None of the sample code could be run on a typical computer; in fact, most of it was pseudo-code lacking a definition of input and output, or even the environment it was supposed to run.

— Steven Wittens

## Graph dependencies

Is maths a directed graph of theorem to theorem? Close enough! But even chapter-level can be helpful:

## Tweaks

- Physicists have a nicer way of marking the variable of integration. Instead of putting

$dx$

at the end, they put it at the start. This saves on brackets and rereading.

$$\int_X$$

## Visuals

It seems insane that *the study of change* is mostly taught without any, y'know, animations.

The limit case of visual mathematics are the lovely proofs without words.

We don't need to endorse any pseudoscience about "learning styles" to think that there are areas of mathematics for which even symbols are not the most efficient delivery.

## Caveats

I'm not claiming that the above are the most important problems with maths teaching. Focussing on mechanical manipulation over insight, and on reproduction rather than creativity, seem like more dire mistakes.

All of academic science is stuck on many of the above, stuck in the 90s. Maybe worst is the stagnation of the conventional paper: static in visuals; never revised unless gross misconduct can be proven; completely decoupled from its justifying evidence and code. Was the last big innovation the hyperlink, 1995? Here are two examples of great post-papers, and a manifesto. (My field, machine learning is unusually tolerant of blog posts, but is still a long way from giving them equal respect, even when it's warranted.)

mathematics is, to a large extent, the invention of better notations

- Feynman

## See also

- Terry Tao on the mathematics of mathematical notation.
- Terry Tao on good notation
- Quantum Country

- Communicating with Interactive Articles

*Credit to John Lapinskas for the induction algorithm.*

{% include lazyload.html %}

# Metabolism is violent

Gavin

2020-09-28

```
{% include phone_img.html %}
```

```
{% assign hammer = "https://archive.org/details/TwilightOfTheIdolsOrHowToPhilosophizeWithAHammer/page"
%} {% assign nintil = "https://nintil.com/longevity" %} {% assign b =
"https://www.sciencedirect.com/science/article/pii/B9780128142530000127"
%} {% assign c = "https://en.wikipedia.org/wiki/Luigi_Cornaro" %} {% assign o =
"https://en.wikipedia.org/wiki/Oxidative_stress" %} {% assign ssc =
"https://slatestarcodex.com/2019/12/12/acc-does-calorie-restriction-slow-
aging/" %}
```

Metabolic stress due to nutrient depletion or nutrient excess triggers a number of adaptive responses to restore dynamic homeostasis and to maintain cellular function.

– Balakrishnan et al (2019)

When there is [an] abundance of nutrients, the signal is to focus on reproduction, while when they are scarce, the cell focuses on reducing the production of, and promoting the repair of, damage.

– Nintil

One of the biggest ideas I've come across this year is that *metabolism is violent*. A “high metabolism” is a source of increased harm as well as increased energy capture. What you want is not *more* metabolic activity, but efficient metabolism, one which maintains you with as little oxidation etc as possible.

Everyone knows that too much food is bad for you; the update is to locate the harm in the body's adaptive response to the excess (its doing more work, expending scarce cellular resources and reducing repair) rather than in the consequences of mere weight gain or first-order toxic effects.

This has a sad practical implication: all but the most careful studies of diet change are heavily confounded by variation in basal metabolic rate (and so on).

The usual idea is that by restricting your diet you live longer.

The sceptical hypothesis is that you have a metabolic setpoint, which determines your diet and your longevity, and which explains most of the correlation.

You need careful experiments to check the latter. There's some evidence that there is a small effect of calorie restriction (evidence mostly from nonhuman models):

the evidence as it stands weakly supports the conclusion that [calorie restriction] modestly extends human life. We expect that an individual engaging in 20-30% CR versus a normative, non-obesogenic diet without malnutrition might enjoy a 10%-20% increase in longevity.

---

Interesting tangent: Nietzsche formed this hypothesis in 1888, right at the beginning of the long diet fad:

No error is more dangerous than that of confusing the cause with the effect: I call it the genuine destruction of reason. Nevertheless, this error can be found in both the oldest and the newest habits of humanity: we even sanctify it and call it ‘religion’ and ‘morality’. It can be found in every single claim formulated by religion and morality; priests and legislators of moral law are the authors of this destruction of reason. Here is an example: everyone has heard of the book in which the famous Cornaro recommends his meagre diet as a recipe for a long and happy – and virtuous – life. This is one of the most widely read books, and several thousand copies are still being printed in England every year. There is no doubt in my mind that few books (except of course the Bible) have wreaked as much havoc, have shortened as many lives as this well-meaning curiosity has done. The reason: confusion of cause and effect. This conscientious Italian thought that his diet was the cause of his longevity: but the preconditions for a long life – an exceptionally slow metabolism and a minimal level of consumption – were in fact the cause of his meagre diet. He was not free to eat either a little or a lot, his frugality was not ‘freely willed’: he got sick when he ate more. But unless you are a carp, it is not only advisable but necessary to have decent meals. Scholars in this day and age, with their rapid consumption of nervous energy, would be destroyed by a regimen like Cornaro’s...

He doesn't deserve too much credit, since he was happy to just assert this and not do the legwork of actually checking it, and since the reality seems to be that diet restriction is just weaker than it looks, rather than useless. But it's a nice illustration of intuitive causal inference.

---

## See also

- Nintil's Longevity FAQ

# Spooky data analysis

Gavin

2021-10-10

On contaminating data by looking at it

Mental contamination occurs when individuals experience feelings of internal dirtiness and distress in the absence of physical contact with a contaminant

Data contamination occurs when datasets experience feelings of internal dirtiness and distress when multiple analyses are run on them or the test set is reused

This is *weird*. It's a set of numbers.

The trick is to realise that it's not the data which is contaminated: the *inference* is.

Specifically, it's circular reasoning. <https://elifesciences.org/articles/48175#bib37>

HARKing is updating twice on the same data

M errors

Data dredging [https://en.wikipedia.org/wiki/Leakage\\_\(machine\\_learning\)](https://en.wikipedia.org/wiki/Leakage_(machine_learning))

People actually use “contamination” to mean noisy data or corrupted data.

# After the boys

Gavin

2020-10-28

Just found out that the climactic line in “Boys of Summer” is not the sublime, tense

“I can’t tell you my love will still be strong / after the boys of summer”

but the flat, deluded, standard

“I can tell you my love will still be strong / after the boys of summer”

The first: Temptation as foregone conclusion, doubt, the acceptance of transience

The second: unjustified assertion

“can’t” makes us demand to know about the boys of summer. Who are they? Why are they taking your boy/girl, or anyway making them doubt? Why is Don Henley’s head getting turned by summer boys? Of course it’s not “after I meet the boys of summer”, “after I fuck the boys of summer”, “after the boys of summer tell me about you”

It’s “post summer”, post youth

I see the boys of summer in their ruin Lay the gold tithings barren,  
Setting no store by harvest, freeze the soils; There in their heat the  
winter floods Of frozen loves they fetch their girls, And drown the  
cargoed apples in their tides.

These boys of light are curdlers in their folly, Sour the boiling honey;  
The jacks of frost they finger in the hives; There in the sun the frigid  
threads Of doubt and dark they feed their nerves; The signal moon  
is zero in their voids.

I see that from these boys shall men of nothing Stature by seedy  
shifting, Or lame the air with leaping from its heats; There from their  
hearts the dogdayed pulse Of love and light bursts in their throats.  
O see the pulse of summer in the ice.

I see you boys of summer in your ruin. Man in his maggot’s barren.  
And boys are full and foreign to the pouch. I am the man your father  
was. We are the sons of flint and pitch. O see the poles are kissing  
as they cross.

# Graphs are cool

Gavin

2020-11-20

{% include graph/links.md %}

If you were to learn one area of maths properly, what should it be?

Depends what you want to do with it. *Send tweet.*

In computer science or machine learning, when people talk about the areas you need, they usually mention calculus, linear algebra, probability theory.<sup>1</sup> These are some of humanity's greatest achievements, and probability can totally change your life, since (outside of mathematics or the wider Set Exercises of school) we have nothing to work with but probable and improbable, priors and data.

What if you're not a technical researcher? What if you just want to get as much clarity as you can, without devoting years of your life to it?

After probability and basic programming, I think there's no area better than graph theory. It is both incredibly intuitive *and* hyper-efficient. It is useful for almost any discrete application: logic, science, society, ... It lets us do lots of things whenever we have "some relations between some objects", i.e. any time we can ditch the continuous.

Obviously this is isn't as abstract as we can go - why have those clunky objects? But it's a nice median.

Getting into the habit of drawing a directed graph is probably the simplest way of thinking better. It takes one minute, and even the qualitative unweighted version will allow you to instantly spot disagreements. I dream of a world where people disagreeing (on Twitter, in debates, in journal letters pages) head to Sketchviz and work out exactly where they're diverging, probably in the relative thickness of two edges. They are astoundingly useful for data-driven science. But they are unbeatable for communication.

## Intuitive

*Proof by inspection.* You can go far in graph theory with visual reasoning.

[TODO: Gif of "every 4-path is self-complementary"]

*Modularity.* You can do lots of things locally, ignoring the overall structure.

It is always nice to be able to reduce some problem to shortest-path or minimum spanning or message-passing or any of graphs' optimal dongles.

## **So many things are graphs**

In some sense anyway, whether it's mathematical equivalence , having a 1-1 mapping (logic), partial capture of structure (groups), or just a useful approximation (society).

Battaglia et al (2018)

### **Sets are graphs**

without edges.

### **Logic is Trees is Graphs**

Any well-formed sentence of logic has a syntax tree, which is a graph.

(Also linked-lists are trees are graphs...)

### **Groups have graphs**

### **Optimisation is shortest-pathing**

Ray tracing and Q-learning and currency arbitrage is graphs. I love this post so much.

See also constraint sat as graph.

### **Graphical models: joint distributions have graphs**

Under very general conditions, joint distributions have graphs. 3

Statistics is one of the hardest things I ever learned. It's just so vast, and even a good grasp of the theory (which almost no-one has) does not prevent 100 completely fatal silent mistakes. Graphs unify the stats zoo.

Many of the classical multivariate probabilistic systems studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics are special cases of the general graphical model formalism – examples include mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models.

- Michael I Jordan

plus PCA, vector quantization, ...

- Old school contingency table stuff

Strictly speaking there isn't a *lot* of graph theory in PGM work. But some graphish algorithms like message passing are still cutting-edge. 4

### Causal inference

Shalizi on causal models.

### Graph neural networks

Convolutions are graphs. (Edge from a node to all neighbours and self.)

The Transformer is a graphnet.

This paper tries to unify many of the discrete neural nets that have sprung up into one framework, and it has an extreme grandeur.

### Massively efficient algos

Many serious databases use trees. i.e. Every formal activity in modern society relies on graphs.

Graph kernels are so fast.

### High theory

Just one instance of it showing up in remarkable mathematics: Ramsey theory for Szemerédi.

Szemerédi's theorem: "any subset of integers of positive (upper) density must necessarily contain arbitrarily long arithmetic progressions."

(This is number theory but the proof flits between that and graphs.)

You get there via Ramsey's theorem: "any finitely coloured, sufficiently large complete graph will contain large monochromatic complete subgraphs."

The path from basic principles (Pigeonhole or Handshake) to huge results like this seems shortest in graph theory. Or maybe that's just my brain.

### Life as graph

Learning the descent of species.

Learning which cells are connected(!)

### Society as graph

PageRank treats the internet as one giant implicit graph.

A lot of the best sociology and epidemiology uses graphs as the core tool.

## Thought as graph

I do not speak, I operate a machine called language. It squeaks and groans but is mine own.

- Dune Messiah

When I write, I am taking the great implicit graph of my thoughts and ripping out a tiny number of particular nodes, and maybe two or three of the many edges. I then serialise these nodes (imposing an artificial order, discarding my weights, idiosyncratic associations, and colour), and have to just hope that you are able to reconstruct some of the original graph from the drips that come through the narrow and impoverished channel of language.

The “don’t call it a notetaking app” notetaking app Roam is designed with this in mind, and it is by far the least unpleasant way of thinking about everything I’ve ever seen.

I am frankly dazzled by how general it is. I haven’t seen any list of the above; I just kept on finding it in new places, it just kept eating objects until I fell in love with it.

Of course you can go much more general in at least three directions: programs, for instance, are much bigger than graphs. But for an *easy* step into rigorous and general ideas they are the winner.

## Drawing quickly and beautifully

Graphviz is kinda painful to use without a live GUI, but Sketchviz pretty much works unless you have very strict spacing in mind.

## See also

- The Fascinating World
- Unifying the Mind

{% include graph/foots.html %}

# Dominic Cummings Report Card (2019-2020)

Gavin

2020-11-14

yes, yes, I mostly don't have enough information to attribute policies to him, and yes this isn't a good model of policymaking even for the most powerful advisors

- Doubling of R&D budget (by 2025)
- Pushed SAGE into an earlier lockdown. That article and their sources try to make this out to be a bad thing, but it would be immoral to do anything else.
- £800m ARPA like
- £300m STEM funding
- Uncapped immigration path for scientists <https://www.gov.uk/government/news/boost-for-uk-science-with-unlimited-visa-offer-to-worlds-brightest-and-best>
- Tutoring in state schools <https://www.theguardian.com/education/2020/jun/17/government-to-fund-private-tutors-for-english-schools>
- Whatever this is <https://www.youtube.com/watch?v=XaN-MbGV4dY>
- Associated data science, superforecasting and evidence-based policy with dickheads, racists, and edgelords. Weirdo scheme may mean Whitehall hiring even more staid in the short run.
- Something something bullying, norms, deception, centralisation of power. Whipping the Treasury probably not a good precedent.
- Hypocrisy probably weakened lockdown norms.
- Not sure how much of the immigration burdens (NHS surcharge etc) are DC.

? No idea if he changed Whitehall at all.

? Hypocrisy. Wasted a huge amount of Johnson's political capital during a public health crisis

# On harmless vice and mindless virtue

Gavin

2020-01-13

People have this naive equivalence between a fantasy and a preference, a voluntary simulation and an increased inclination to

Let's imagine that porn industry was safe and there were no survival sex workers.  
Is watching porn unvirtuous independent

What about shooting people in GTA?

# What is the philosophy of mainstream economics?

Gavin

2012-03-13

Disclaimer:

*I was an angry young man. One of the targets of my ire was academic economics. Until the spell of Mirowski and the heterodoxy, I found myself smarter than thousands of people who knew more than me and worked harder than me.*

*Part of this was double standards: “social science must be perfect to be legitimate” - where social theory, an interpretation of society, needs no grounding besides eloquence and grandeur.*

We can only conclude that economics, as studied in our universities, is the astrology of the Machine Age... In science the final arbiter is not the self-evidence of the initial statement, nor the facade of flawless logic that conceals it. A scientific law embodies a recipe for doing something, and its final validation rests in the domain of action.

- Lancelot Hogben (1936)

Economic theory is mathematical analysis. Everything else is just pictures and talk.

- Robert Lucas

My problem with economics stretches to every major question of scientific methodology: from defining the field; to quantifying its domain; the relationship between a science's methods and object; examining its fundamental principles; to the role of simplifying assumptions; a large family of issues regarding evidence's limited role ("econometric methodology"); the supposed distinction between positive, normative, and applied uses of theory; the role of normative concerns in science; the rhetoric of economists; the role of conformism or pluralism; to in what sense economics even has scientific status. Though I do not have space to address all of these here.

Call the dominant paradigm *purist economics* (PE).

Call the dominant theories in the paradigm mainstream economics. The more common name is ‘neoclassical’ economics, but this obscures its object in two respects – firstly historically (because the original ‘neoclassicals’ were C19th economists, and were quite different from our contemporaries) and, secondly, because the term is made to cover two orders of explanation (‘neoclassical economics’ being used for both the dominant theories, and the meta-theory that orchestrates them). Problems with economics fall into three groups: the technical, the philosophical and the institutional. These sets are not independent, and might be characterised by the causal chain:

Fig.1. quasi-Marxist model of the methodological problem of economics And though simplistic, there is something to this. I thus include a section on explicitly institutional factors. I will not discuss technical failures in this essay, mainly because of the considerable insider knowledge they require. (But lack of predictive power is the key one, and that is conspicuous to non-technicians – as in the Great Recession of 2008-Present.)

One respected economic methodologist, Uskali Maki, has a warning for people like me: “Economics is a very complex subject, and any given account of it will only highlight some of its limited aspects, serving only limited purposes. Overgeneralized and oversimplified accounts abound, and they are just that: overgeneralized and oversimplified.”

Say that purist economics is a cluster centred on the following beliefs:

1. Love of precision, i.e. of quantitative results in a Newtonian mould.
2. Love of abstraction, born of the belief that maximising abstraction will give the model maximum relevance.
3. Monomathematical. That mathematical modeling is the defining tool of economics –“I came to the position that mathematical analysis is not one of many ways of doing economic theory: It is the only way. Economic theory is mathematical analysis. Everything else is just pictures and talk.”
4. Methodological individualism. (2) compels us to focus on idealised unitary objects: analysis must deal with discrete units if certain mathematical tools are to begin to work.
5. Micro reductivism. (3) entails a quixotic search for the ‘microfoundations’ of macro-phenomena, because otherwise they cannot be said to be theoretically ‘real’.
6. That Lionel Robbins’ 1932 definition of the field is the salient one: “Economics is the science which studies human behavior as a relationship between given ends and scarce means which have alternative uses.”
7. Rational Choice theory. That utility theory is the basis for modelling human behavior. This squashes all human experience to one dimension: utility, all you can get. That economic rationality is and only is utility maximisation.

In supplying unfalsifiable behavioural premises, the above combine to make PE largely deductivist.

8. One-dimensional men. (6) dictates that explanation of any given human action should be in terms of prudence (i.e. greed). Put technically: “an agent’s unbounded monotonic utility maximisation”. In fact: economic explanation is to demonstrate that agent A optimises some variable x.
9. Platonic belief in equilibrium. That explanation of a given economic phenomenon is simply the identification of its ‘equilibrium’ (that idealised point where forces co-ordinate and stabilise the system). Equilibria are taken to be ubiquitous. Price determination is the primary product of analysis.
10. Heroic assumptions. The precise quantitative standards set by (1), (3) and (8) imply very restrictive and unrealistic assumptions about human behaviour, typically perfect information (that agents are all-knowing), perfect rationality (narrowly construed) and unbounded monotonic utility maximisation (that our wants are unlimited).
11. Imperial scope. From (6); that economic analysis can be rightly applied to any human behaviour. All decisions have an internal logic based on (7)&(9).
12. Methodological instrumentalism. Since they require patently false stipulations, the above heavily motivates the position that only the power of one’s results matter in analysis, not the starting premises.<sup>7</sup>
13. 1920s-grade Positivism. That economics can be neatly split into positive and normative methods - and that the former can be value-free. That politics and economics are separate.
14. The futility thesis: Philip Mirowski notes that NE theories fit a pattern: most end with “some assertion that the world is so structured and interlinked that anything one might wish to accomplish will be offset ... resulting in a return to the original situation...” (see (9).)
15. *Exemplars*. Arrow (1951) and Friedman (1968).

## Philosophical failures

### PE’s questionable scientificity

Edward Lazear, a powerful policy-maker, gives PE’s understanding of itself:

Economics is not only a science, it is a genuine science. Like the physical sciences, economics uses a methodology that produces refutable implications and tests these implications using solid statistical techniques...

Let us see.

Here is a general schema for evaluating a science: 1. What kind of science is it? 2. Are its goals scientific? 3. Are its methods scientific? a. Are its foundations empirically founded? b. Does it proceed empirically? c. Is it falsifiable? d. Is it consilient with the natural sciences? 4. Are its conclusions scientific? a. Does it find laws? b. Is it predictively powerful? c. Is it explanatory? d. Does it make progress?

And – pace Lazear – for PE the answer to each of (2)-(4) is ‘no’. (Note that these questions address what the goals, methods and conclusions of the field have been – but it is vital to ask in addition what they ought to be.) First: What kind of science is economics? We can answer this by either studying economists’ practice or by analysing their objects. Let us try to characterise a typical macroeconomic phenomenon: it has intentionality; it is what Hilary Putnam calls “thick” (i.e. inextricably evaluated as it is described); it is essentially contingent and dynamic; it is overdetermined; formally indeterminate; it is reflexive (i.e. can react to theory); and controlled experimentation is often practically and ethically impossible. Figure 2 is my attempt to capture the problem with calling PE a ‘genuine’ (that is, hard and natural-like) science. I use ‘science’ in the older, broader sense of ‘a systematic way of understanding empirical events’:

The error is complex: purists speak as if economics were equivalent to natural science when it is more often a merely formal science (living as it does in a corner of pure mathematics) – and when it should be a policy science.

The intuitive split they rely on is between ‘fact-based’ science and ‘value-based’ ethics; effectively an attempt to apply positivism to humans. This split is contentious even in natural sciences, but simply inadmissible in social ones. Facts are interpreted, and there are dimensions of intentional objects (like actions, beliefs, desires and utility) that cannot be properly understood at all without reference to psychology and phenomenology. After Gunnar Myrdal, I argue that social science must aim at *verstehen* (participatory understanding) as well as *begriifen* (understanding of causal mechanisms). Analysing the origins of PE, Philip Mirowski finds that it reconstructed itself in the image of physics, taking up Newtonian goals (quantitative results, maximum precision, predictive power, *begriifen*).

In the name of objective economics, the primogenitors of purist economics – most significantly Leon Walras, Vilfredo Pareto, and Lionel Robbins – intentionally depychologised, depoliticised, denormativised economics, eventually arguing that the market wholly sidestepped issues of justice. But wrestling with the problems of a ‘human physics’ led to PE resembling pure mathematics far more than physics. In practice, PE is a formal science with pretensions to synthetic knowledge. An ideal economics would be far closer to law or history than to physics. So, regarding schema question (2): its goals are indeed scientific – but nonetheless unjustifiable and unattainable, in context.

The practical problem with PE is that its claim to hard science serves to falsely

elevate economists' policy advice, and immunise the field to criticism from non-economists (and even to economists who are not monomathematical). Warren Samuels lists the many things that PE actually is: "... economics could be seen as science, as political and moral philosophy, as ideological self-projection by the people of the Euro-American nations and their way of earning a living, as both derived from & generally reinforcing the existing structure of power, privilege & so on."

This leads us on to PE's problems attempt to avoid normativity.

### The 'Value-free' myth

Here is a reconstruction of a common intuition in PE, and an application of it:

P1. The world is value-pluralist. (People disagree on what matters.) P2. So to take any normative view is inevitably to make yourself unobjective. (To be universal, one must be value-neutral.) P3. Real sciences are objective; so real sciences are value-neutral. P4. Economics is scientific. P5. So economics must be value-neutral. (3&4) P6. So economics can yield objective facts about economies. P7. The market aggregates the values of all individuals into a price. P8. Price is not normative because it is just a neutral channel for individuals to establish a group value. P9. So market price is an objective value (an intersubjective summary of participants' values). Price circumvents value judgments. C. So by utilising market data, economics can give objective policy advice. (5&9)

We have already argued against P4. But the implicit normativity of economics goes beyond the intentionality of its objects: the limits of empiricism are more keenly felt out in social science. Some general processes that militate against this neat argument:

- *Naturalisation of contingencies.* John Lanchester summarises one of these limits well: "Empiricism, because it takes its evidence from the existing order of things, is inherently prone to accepting as realities things that are merely evidence of underlying biases and ideological pressures. Empiricism, for Marx, will always confirm the status quo."
- The necessity of interpretation. And Lanchester continues into a second issue, the theory-ladenness and interpreted nature of all facts: "[Marx] would have particularly disliked the modern tendency to argue from 'facts', as if those facts were neutral chunks of reality, free of the watermarks of history and interpretation and ideological bias and of the circumstances of their own production."
- The normativity of definition and of emphasis: The tools decide the issues that can be addressed. As Samuels put it, "Phenomena do not come with the label 'economic' attached to them."
- The Duhem-Quine Thesis is particularly dangerous in economics, since

testing takes place in the presence of very dubious auxiliaries. Hausman: “if one is unable to place much confidence in the other premises needed to derive a prediction P from an hypothesis H, then there is a serious practical problem. Indeed it becomes almost impossible to learn from experience. This is the situation in economics”

- Performativity of economics. The above all hold in, for example, physics. But there is little scope there for dangerous smuggled normativity and performativity found in economics. Economic propositions are performative in the sense that their subject matter covers not simple mechanistic events, but intentional objects that are often blends of belief and event – e.g. “I am happy” – and of illusion and reality – e.g. “Capitalism is natural”. Also in the sense that successful theory can affect economic processes, because economic actors can read it and alter their behaviour. (In Donald MacKenzie’s phrase, theory is “an engine, not a camera”.)

The above combine to make empirical data generally insufficient to strongly guide researchers in theory-choice. Another conclusion we can draw is that economies cannot be separated from the political and social system within which they are embedded; neither can economics. The attempt to make economics objective was well-motivated – but misguided. The subtraction of ethical concerns from the field led to worse policy and falser claims.

### **Formalism -> Omitted variables**

The overwhelming focus on mathematical modelling leads to distortion in PE’s subjects. (Again: one’s tools decide the issues that one can address.) PE proceeds by marginalising or omitting what it cannot formally represent. We have seen the results of omitting ethics from analysis, but the omissions actually amount to far more – most of what actually determines economic life: personality (“heterogenous decision rules”); dynamic social contexts; time; path-dependency; competition; preference management; processes of resource distribution; ideas (entrepreneurship); transaction costs; profit; and vital non-economic coordination, like law. If and when this process is conceded, the word ‘exogenous’ (meaning simply ‘external to the model’) is used to avoid explaining the variable. In applied science, there is a trade-off between the relevance and precision of one’s model. PE denies this trade-off, and ends up maximising precision at the expense of their model’s realism: abstraction, in the end, guarantees applicability only at the cost of vacuity.    ##### Ahistoricism

Perhaps 90% of the world’s economics departments operate within PE; almost none of these teach history as standard, and an increasing number have no historians of economic thought. This indifference to history (as pedagogical or theoretical source) actually serves to keep PE dominant. Daniel Garber gives a methodological argument in favour of philosophers doing philosophy of history, one which transfers well to our case. [Economists] should study history

not because the past is a source of arguments and positions that may be helpful in

current debates - though it is that - but for a more important reason... History gives us fresh views of what philosophy has done and so can do - of what it was and so can be. And this is what is needed by problem-solvers today.

Claim: Study of history is intrinsically harmful to dogmatism, because it contextualises one's work and paradigm.

### **Institutional failures**

Economics gets called the 'queen of the social sciences', with General Equilibrium its 'crown jewel'. This holds – but only if we think of PE as Lewis Carroll's Queen of Hearts – i.e. jealous, duplicitous, and overbearing – and GE as a rhinestone. I do not think paradigms are avoidable – but, as with any monopoly, there are risks attendant on it. Some negative consequences of economics' paradigmatisation:

#### **Internal tyranny**

There is a wilful incommensurability in PE: "Sneering is the obstacle to conversation in economics. The Chicagoans sneer at the Marxists, the Marxists sneer at the Neoclassicals, the Neoclassicals sneer at the Austrians, and the Austrians sneer at the Chicagoans.  $C < M < N < A < C$ ." I have deemed PE a paradigm, so in one sense we cannot expect there to be anything but strict enforcement of methodology. But, on the model of biodiversity, here is a strategic argument for scientific pluralism: 1. Science is provisional; it can be replaced at any time. (This caution is motivated by Larry Laudan's 'pessimistic meta-induction': the realisation that all past theories have been superceded - so if we are to infer anything from the past, it is the likelihood of our theory's future refutation.) 2. Any given science always faces anomalies, mysteries, tensions, and critique. 3. If a whole community is of the same research programme, and meets a persistent anomaly, then the whole science is blocked and undermined by it. 4. Pluralism gives the field alternative paths and approaches, increasing the chance of incorporating anomaly. 5. Methodological pluralism is the enlightened option (since (1), (2) and (4)).

And a further argument from efficiency can be made: a very great part of the field is concerned with "static equilibrium conditions", that is, on finding one special list of prices at which all the products of an economy will be bought and sold at a particular point in time. To use the language of economics: in chasing over-idealised and inapplicable research programmes, PE results in gross misallocation of scarce intellectual resources.

#### **Insularity**

One response to the demand for pluralism is: "Why must every academic try to speak about everything? Why can't the other social sciences pick up the touchy-feely bits of the work?"

This is a fine point but for 1) the intellectual imperialism of economics (the attitude and method described by our cluster criteria has spread to other fields, and occluded human issues there too). Also 2) the performativity of economic theories: economic education alters the worldview of economists, many of whom go on to do ethically-loaded policy work with a false sense of objectivity. And finally: 3) economics has been too closed-off from other departments to make that question's implied division of mental labour possible: citation analysis (volume of non-economics papers quoted in economics) shows it to be the most isolated large field outside of pure mathematics. (Thomson Reuters (2012) Web of Knowledge database, <http://isiknowledge.com/>) No less a purist than Hayek knew the danger of this insularity: "...nobody can be a great economist who is only an economist - and I am even tempted to add that the economist who is only an economist is likely to become a nuisance if not a positive danger." Sociologists call the process of making a new field distinct from others 'boundary work': it involves inventing a jargon and asserting one's scientific credentials.

The strategic argument for pluralism given above maps easily on to this. Hausman puts it drolly: "The only legitimate reason to dismiss all work of other social scientists as of little interest would be if the separate science of economics were a smashing success."

#### Applied cacophony

Despite the methodological enforcement outlined above, even canon PE economists fail to agree on almost every actual real-world application of the theory. The cause of this has been euphemised as "the art of economics" (as if this were not truly a failure of the science at all). But this raises the question: is PE in fact a paradigm? In Kuhn's work the term implies consensus on method and on most substantives. So this could be a further block to PE's claim to hard-scientific status: it lacks the 'univocity' of a Kuhnian mature science. If so, economics suffers the disadvantages of paradigmatisation without the primary benefits: power and univocity.

Anti-philosophical methodology "...without exception, we all share philosophical background assumptions and presuppositions. The penalty of not doing philosophy isn't to transcend it, but simply to give bad philosophical arguments a free pass." - David Pearce

Methodology is a science's site of self-reflection and self-criticism. The above problems are as universal and intractable as they are because PE's methodology disparages work on methodology: if it is considered at all, it is deemed a "misallocation of scarce intellectual resources". Economists can literally merely cite Friedman's 1953 paper and go on with an inconsistent faux-positivist faux-Popperianism. This attitude to methodology could be justified, if the paradigm had demonstrated its success and scientific rigour – but, to put it mildly: this is not the case.

In conclusion: PE's methodology has been seen to be an opportunistic blend of positivism, misused Lakatosian ideas, and inapplicable natural-science method-

ology. (Most of my conclusions apply to microeconomics more than macroeconomics, but some processes are shared.) Figure 3 is my composite of effects:

Fig.3. A slightly more sophisticated historical model.

As a result of these(largely unintended) features, the paradigm could not shift despite sustained criticism, the accumulation of serious anomalies, and the availability of multiple well-developed alternatives. It remains to be seen if that titanic anomaly, the current 'Great Recession', will overcome purist economics' strong inertia.

## See also

- <https://www.overcomingbias.com/2020/08/common-econ-critiques.html>

### <h3>Bibliography</h3>

<div>

- Arrow, Kenneth (1951), 'An Extension of the Basic Theorems of Classical Welfare Economics'
- Colander, David (2004) 'Economics as an Ideologically Challenged Science', working paper,
- Colander, David (2007) 'Pluralism and Heterodox Economics', working paper, <http://sandcat.net>
- Colander et al (2009) 'The Financial Crisis and the Systemic Failure of Academic Economics'
- Dow, Sheila (1997), 'Mainstream economic methodology', Cambridge Journal of Economics vol
- Friedman, Milton (1968), 'The Role of Monetary Policy', American Economic Review 58, No.1
- Friedman, Milton (1953) "The Methodology of Positive Economics" in his Essays In Positive Economics
- Hausman, Daniel (1992), The Inexact and Separate Science of Economics, (Cambridge; Cambridge University Press)
- Hayes, Christopher (2007), 'Hip Heterodoxy', The Nation, <http://www.thenation.com/article/hip-heterodoxy>
- Lanchester, John (2012) 'Marx at 193', London Review of Books, <http://www.lrb.co.uk/v34/n06>
- Latsis, Spiro (1972), 'Situational Determinism in Economics', British Journal for the Philosophy of Science 23, No. 91
- Maki, Uskali (2008) 'Realism from the 'lands of Kaleva': an interview with Uskali Mäki', <http://www.uskali.org/reviews/maki.html>
- Merton, Robert K (1957) Social Theory and Social Structure (Glencoe, IL; Free Press; 1957)
- Mirowski, Philip (1989), More Heat Than Light: Economics as Social Physics, Physics as Natural Science (London; MIT Press; 1989)
- Mirowski, Philip (2004) The Effortless Economy of Science? (Minion; Duke University; 2004)
- Myrdal, Gunnar (1954) 'Implicit Values in Economics' in The Political Element in the Development of Economics (London; Hutchinson Educational; 1954)
- Putnam, Hilary (2011) The End of Value-Free Economics (Oxon; Routledge; 2011)
- Stiglitz, Joseph (2010), 'Needed: new economic paradigm' <http://goo.gl/yWqV3>

</div>

### <h3>Maturity or co-option</h3>

<div>

The original title of this essay, when I wrote it in 2012, was "What is wrong with the current economic paradigm?"

# A decomposition of decompositions

Gavin

2015-09-17

Age is at least five different things which we currently, sensibly, treat the same. (We do this by using just one integer, ‘years since birth’, as its only measure.) What things is age?

Periodisation. A person’s place in history, extremely well covered by date of birth. Through DOB we get a sense of what cluster of opinions they will probably hold.

Biological age. A person’s senescence. The age-integer is also used a proxy for how much help a person needs or deserves, and how much production you can expect from someone, with 65 years an arbitrary threshold in most of the developing world. Philosophically, it would make a lot of sense to collapse old-age welfare into disability welfare, since old age is disability, and since both resource allocations seek the amelioration of a difficult life. But, politically, this would be a bad move for the old, since it’s pretty easy to slash disability spending but (currently-disbursing) pensions are heavily guarded.

Total subjective time. How much have they been through, consciously? This measure is not respected yet; for instance, we call people who wake from long comas by the age indicated by their date of birth, and expect corresponding behaviour from them. What does dementia to this variable - do forgotten experiences not count towards one’s subjective age? does forgetting make you ‘younger’?

Social status allocated. Much of history was gerontocratic: you served your time and earned power just by being old. This pressure (which led to e.g. polygamy for the old élites) is at odds with the presumable motive of judging people by age type (2): as proxy for reproductive fitness. Western culture has probably overcompensated in the other direction by now.

Wisdom or maturity. We even try to use the age-integer as a measure of profoundness and credibility, probably as a result of (4). We call wise young people ‘old souls’. When staying alive was a hard thing to do, (2) was informative.

At the moment, the age-integer carries a lot of mutual information about these 5 things. But we can expect this to decline; technology is beginning to unpick the senses. (1) and (2) are already quite divergent: people with the same date

of birth vary widely by metabolic and mental integrity. Genetic engineering could make this a chasm: think of the social upheaval of a 100 year old CEO, Olympean; a cryonics survivor with 200 years between their DOB origin and the apparent wear on their body; living people who remember the days when women had to drag around new people, often unto death. Memory enhancements could affect (3), the phenomenology of age in hard to conceive ways. (Some fictional evidence here from a master of the barely conceivable). Much later, in space, time dilation and [whatever hibernation method sticks] could make (1), (2) and (3) diverge complexly; when, in Interstellar, the doctor tells Cooper he looks good for being 127 years old, he is saying something importantly false, because (3) Cooper did not experience, and (2) his body does not wear 80 of those years.

Some of you will be thinking ‘Huh! The age-integer sucks. Let’s not use numbers to categorise people’. On the contrary! we just need four more good ones.

*This is surely not novel, but it was original, so I’m recording it as an early (2015) solid piece of conceptual analysis.*

# Irony, sincerity, nostalgia, neoteny

Gavin

2021-01-09

```
{% assign ger = "https://twitter.com/geraldstratfor3?s=09" %} {% assign nine = "https://youtu.be/9wFwPh-KbEY" %} {% assign auer = "https://theamericanreader.com/jenesuispasliberal-entering-the-quagmire-of-online-leftism/" %}
```

Friend 1: Do you know any writing on the recent trend of millennials/zoomers going for ‘earnest’ or ‘wholesome’ content

Friend 2: We live lonely, disconnected, meaningless lives in a broken society on a dying planet. If *Avatar: The Last Airbender* or videos of people rescuing ducks brings a moment of emotional well-being into someone’s life, so be it. I also think it’s for the same reason that we’re currently obsessed with the end of the world and can’t get enough zombie movies and games about nuclear war.

Friend 1: I guess a better question would be: “Is youth culture inherently sardonic and ironical, and do people deep down resent that.” Why do you think our culture is so nostalgic now?

Friend 2: I think it’s a misstep to assume that the Millennial/Zoomer desire for wholesomeness is ironic. I think that, even if it was initially ironic (out of a sort of defensive desire to be ‘cool’) or whatever, a lot of it is entirely sincere now. Personally, about 3 years ago, I watched all of *The Last Airbender* while I was recovering from a really bad panic attack and it genuinely made me feel a lot better.

Friend 1: No sorry my point was the default mood of our times is irony, and people’s desire for earnest things is genuine and sincere, I didn’t make that clear.

Friend 2: The easy answers are that people are trying to go back to “a better time”, but considering that most of the people celebrating 1980s culture weren’t born until at least the 1990s, it can’t be only that. I personally think that a lot of it is to do with the 80s being enormously optimistic (in an aesthetic sense). 80s design was all about the future - sharp angles, bright colours, an obsession with technology. But, more than that, it’s not only the 80s - the 90s are

coming back now as well. Surely it can be argued that a desire for sincerity and earnestness is exactly because the default mood of our time is ironic? Is it too simple that one is simply a response to the other?

Friend 1: My thoughts too. I think the big cultural change of the last twenty years is people extending their adolescence way into the late 20s and 30s. And the rise of ‘don’t know how to adult’ lingo. Maybe the nostalgic thing is part of that. Like I think you could make a fair case for rumours by Fleetwood mac being one of the most universally popular albums with under 40s atm. By which I mean amongst middle-class people in Canada etc. (I’m aware that if you live in raqqa you probably have other cultural shifts that come to mind when musing on the last 20 years.)

Friend 2: I think that extended adolescent thing has more to do with technology than people realise. Technology is literally changing the way that society expects people to work, socialise, present, behave etc faster than they can adapt to it. In response, I think a lot of people defer committing to ‘a life’.

---

Me:

I am unpersuaded that the impulse to sincerity is new (that is, less than twenty years old), it’s just more visible now.

One good keyword is “the meaning crisis”: i.e. the absence of satisfying total objective systems that make you feel like you understand your role, can measure your success.

No one runs surveys on this so I’ll use famous manifestos as a proxy since I just need to show that there are examples over time:

- 1770: Young Werther fever. Landmark for when the culture (in German youth) was way *too* earnest.
- 1890: Nietzsche. Prediction that people will get ironic and weird after religion. . .
- 1985: Kitsch. “New Sincerity” (DFW obvs, but moreso people like Albom and Sedaris)
- 1999: Dogme 95, Stuckism
- 2005: EMO! An absolute powerhouse, maybe a quarter of all teens in some way.
- Now: Wholesome memes, mumblecore, cottagecore

Is youth culture inherently sardonic and ironical?

Pro: Many teenagers are in constant low-level pain (acne, growing, sleep deprivation, boredom, uncertainty, self-consciousness, depression) and pain makes us sharp and anti.

Con: I think teens just need to do the opposite of what's around them. Irony is pervasive now, so (some) go anti-irony. But sarcasm is versatile, so you can still be sarcastic about sarcastic people; you can negate from inside.

There's also a hybrid pap of sincerity and irony which is maybe more common than sincerity: consider any Disney / Marvel / etc product: almost every line of dialogue will be sassy, and parents are more often than not the bogeymen, but underneath the work is black-and-white morality tale, with friendship, folk deontology ("we don't trade lives"), and vague but ironclad humanism.

---

Allow me some silly theorism, to drag in abstruse academic bullshit to this discussion of silly pop culture products. Some people blame intellectuals for irony. There's a huge cottage industry of secular prophets denouncing irony. They call it "cultural Marxism" or whatever the almost-made-up bogeyman is. Think Jordan Peterson, James Lindsay, and many more unsavoury characters.

I think they are literally twenty years too late. Twenty years ago, if you wanted to be an intellectual (which stupidly means: "person who thinks about society"), you had to pay some lip service to the French Theorists, or set out in explicit opposition. But Theory is dying. In its place are a new wing of anti-pomo dogmatists with nothing in common with them except radicalism and jargon. The rise of cancelling and the political domination of inner life *could not happen* in a truly ironic worldview. Instead we get the Marvel version of irony: sassy Puritanism.

Is youth culture inherently sardonic and ironical, and do people deep  
do...

It's inherently about looking cool, and among emotionally incompetent people in pain, a perennial shortcut to coolness is negativity, detachment, and seeming not to care. But both Tony Stark and Captain America are cool: sincerity has always been an option in youth culture

Why do you think our culture is so nostalgic now?

I buy Friend 2's bit: nostalgia is a way of fighting loneliness. If there are now millions of subcultures, if it is now harder to interact with a random peer because the options spread us all out so much, then pop culture and remakes allow us to go back to when there was a shared milieu.

This reinforces neoteny, our twenty year adolescence: you can easily go back to the old if you're not very emotionally distant from it.

Ofc adults in the past were equally bewildered, if not more bewildered than us. But the taboo against admitting it disappeared some time in the last 30 years. Slackers and losers gained some ground. Oh and ofc nerds massively gained status, and we are highly childish in two senses: social incompetence and open enthusiasm

---

We live lonely, disconnected, meaningless lives in a broken society  
on a...

Another reason for nostalgia, fleeing backward, is if you believe the present is broken and not going to improve. My big prediction is that this reflexive pessimism is going to drop this decade. This year has seen 4 or 5 giant tech breakthroughs, and there are more coming soon. I also think the exaggerated timelines and extent of the climate emergency will lead to a fall in doomy greens.

Progress -> optimism -> sincerity -> disappointment -> backlash ->

It'll come back of course. And pessimism is more robust to facts than optimism, so it'll never go away entirely.

When was the last time "The future is going to be fun" was a dominant view? 60s?

---

I'm on a higher level of bloody-minded contrariness than your average bear

1. the ironic eye rollers
2. the new new sincerity wholesomes
3. the modernists (neolibs)
4. the postmodernists
5. the vulgar anti-postmodernists
6. the illiberal left and right

So despite being an edgy stemlord I will be reading Foucault and Bourdieu this year.

---

If this discussion was social science rather than mere criticism, we'd have to unpick the several dimensions being squashed:

1. Positivity / negativity
2. Sincerity / irony
3. Concreteness / theory
4. Ethics / Structure
5. Solidarity / suspicion
6. Realism / relativism
7. Pleasure / Maturity

with culture war left/right stuff jittering everything a bit across all these axes.

The only reason this post isn't useless, that it isn't fatal to squash them, is that many people allow themselves to be psychically herded into a flat 1D projection of real political/ethical//existential space.

Auerbach: trust x agency.

---

Ok I need to shut up but:

Irony is good because it's sceptical, and we run a permanent yawning deficit of scepticism as a civ.

Irony is bad because it's deadening, it interposes itself between you and direct enjoyment, direct communication, the ding-an-sich or even ding-für-uns.

Sincerity is dangerous because it *must* be credulous, it must take preferences, experiences, naturalness for granted, at least in the moment. Chronic sincerity invites manipulation, disinformation, and so on. As always, we must work out how to maintain both.

One solution is to be sincerely truth-seeking, honestly critical. DFW is a model, or Rorty's "*yes they're a eurocentric fiction but I still insist on human rights*". These stances are unavailable to most people, because they don't feel very good: they're a tightrope act of negative capability. Almost no one can apparently do this. I have to watch myself or else I fall into a culture war faction, or an anti x mood, or self-satisfied technocrat superiority.

### See also

- Irving, A constructive critique of Sapiens and Homo Deus
- Against the Culture
- Notes on Infinite Jest
- Strangers Drowning
- Blindsight

# hope dies last

Gavin

2020-12-02

```
{% assign dying = "https://en.wikipedia.org/wiki/Dying_Earth_genre" %} {%  
assign f = "https://tvtropes.org/pmwiki/pmwiki.php/Main/ThePowerOfFriendship"  
%} {% assign m = "https://en.wikipedia.org/wiki/Manichaeism" %} {%  
assign rs = "https://www.polygon.com/tv/2018/9/3/17806570/adventure-  
time-finale-finn-bubblegum-marceline-adam-muto-interview" %} {% assign  
time = "https://www.youtube.com/watch?v=Xr53S9vIbCE" %} {% assign  
ac = "https://en.wikipedia.org/wiki/Anton_Chigurh" %} {% assign stag  
= "https://adventuretime.fandom.com/wiki/Stag" %} {% assign matt =  
"https://adventuretime.fandom.com/wiki/Matthew" %}
```

*Adventure Time* is a cartoon about the fantastical anime/D&D quests of a boy and his anthropomorphic dog. Superficially, it is joyous: filled with treasure, candy, fart jokes, dayglo idiots, new slang, and dance parties.

But the joy in it is *post-post-apocalyptic*: the show is undermined and deepened by a dark frame - the adventures take place in a ruined Earth, with ~all of the adults gone. It is an unusually good depiction of nihilism, trauma, the ‘meaning crisis’, being neuroatypical, the contingency of self, virtue, success, love, and even existential risk.

Without reading between the lines, the show is just normal Cartoon Network Power of Friendship fare. But the real theme of the show is how to be happy in a hostile, finite, godless world. Probably only about 50 of the 280 episodes are about this, but it’s the rich part.

DISCLAIMER: There are about 700 characters in the show. With this many degrees of freedom it’s possible to support most weird readings by being selective.

## Glory passes

Manichaean religion viewed the world as a finite war between light and dark - with light constantly leaking out of the world, unto eternal night. So too in *Adventure Time*: as the series goes, more and more of the heroes, gods, and stabilising forces in Ooo are removed:

- The greatest hero Billy retires, broken. Then corrupted and used by xrisk
- The god of Mars, GrobGobGlobGrod is killed.

- Prismo - an even more godlike God - is killed.\*
- Matthew, a hive mind who claims that he will restore the world after the next apocalypse, is murdered for no particular reason.

Against this steady loss of checks and balances, villains disappear too. Because they die, or because they are aggressively humanised.

- Ice King -> heroic Simon,
- Lich -> Sweet Pea
- Darren
- etc

\* Some of this post was written before the last few seasons brought people back and generally pushed against entropy.

## Injustice

The least childish thing about the show is the repeated instances of unpunished evil and unrewarded virtue.

- Magicman, a sort of camp Anton Chigurh, does many heinous things, including summoning a world-eating monster. He suffers somewhat, but recovers and goes on being heinous.
- Root Beer Guy, a blameless minor character given one very heartfelt episode, is brutally murdered during a siege of the Candy Kingdom. All others who die in this episode are revived, but not him.
- The terrible Stag, who briefly enslaves an entire country in order to devour them, is transformed into a telescope (justice I guess), but is later put back.

## Martin: the fragility of self, the accident of virtue

Martin is Finn's estranged dad. He is introduced as a horrible rogue with no paternal feeling whatsoever. Later we realise that he was actually separated from Finn while heroically defending the boy, and got brain damage.

Fans dislike the brain damage idea, since it feels to them to rob his actions of the evil and arrogance they want to ascribe him. It's true that cheap writers use brain damage as a way of getting out of plot corners. But this instance is neither cheap nor dissatisfying:

Martin the husband tried his best, was even heroic, and still ended up failing his family - and worse, ending up with them thinking he was a villain; and worse, actually having his personality altered to confirm that impression.

The point is that the self is fragile; virtue and vice are partly happy or unhappy accidents; and yes the mask can eat into the face and make you into what you do.

He was a rogue and a cheapskate before Minerva, but he is much much worse than this after the head injury. You can retcon this as his defence mechanism from shame, but I find the neurological explanation simpler, more disturbing, and satisfying.

### What do we know?

- We have one bit of concept art showing him after the Guardian fight with a head wound.
- He vaguely remembers Finn and sits with him in a young dad way. In my telling, this is unconscious muscle memory.

### Lemonhope dies last

The most startling episode in the whole series is Lemonhope II. A thousand years after the events of the main series, Lemonhope wanders through an empty wasteland, passing abandoned cities and fallen landmarks. We see no other life. Then alone he dies.

The timeline it implies:

1. Our Earth
2. Apocalypse 1, the Great Mushroom War
3. Shoko & Tiger
4. Finn & Jake
5. Apocalypse 2
6. Shermy & Beth
7. The death of Lemonhope, the last adventurer

That is, this cartoon depicts the total death of its own world.

### Or maybe everything stays

Understandably, the finale veers away from this, instead emphasising endless cycles of death and new life. It hints that Lemonhope II isn't the final word on Ooo, that it goes on: "people just keep living their lives". There are more and more apocalypses, and people recover and get used to it each time. "Everything stays, but it still changes"

I don't mind this alternative; unlike the reading of Martin that paints him as *just* a wilful liar, or *just* the result of deeply repressed shame, the eternal adventure is at least an ethos.

### Fandom is forever

The final song, "Time Adventure" has a lot going on. First: 4-dimensionalism about time.

*Time is an illusion that helps things make senseSo we're always living  
in the present tense.It seems unforgiving when a good thing endsBut  
you and I will always be back then.*

It's common to deny that good things were good, e.g. following a breakup. Like Plato: "not real if it doesn't last forever". But on plausible views of time (growing block or eternalism), *the value still exists after it is over*: nothing subsequent can ever touch it. The universe's heat death (the end of this show) is bad, if it is bad, because it stops us having more value, not because it nullifies past value / meaning. I find this incredibly helpful to steer through life.

Second: "Time Adventure" has the characters directly address us, the audience.

*If there was some amazing force outside of timeto take us back to  
where we were And hang each moment up like pictures on the wall  
Inside a billion tiny frames so that we can see it all, all, allIt would  
look like:Will happen, happening happened...*

That force is you, e.g. watching favourite episodes out of sequence, e.g. writing long strange rants about headcanon. Possibilist reference, if you like. Whatever its internal fate, Ooo is immortal already because we're outside their time. It's not gone until we're gone. It cushions the cancellation of a beloved show with a sermon on the serenity of a good philosophy of time and reference.

# The art of procrastination

Gavin

2018-08-24

structured procrastinator: a person who gets a lot done by [consciously] not doing other ['important'] things. . . structured procrastination requires a certain amount of self-deception, since one is in effect constantly perpetrating a pyramid scheme on oneself. Exactly. One needs to be able to recognize and commit oneself to tasks with inflated importance and unreal deadlines, while making oneself feel that they are important and urgent. This is not a problem, because virtually all procrastinators have excellent self-deceptive skills also. And what could be more noble than using one character flaw to offset the bad effects of another?"

Previously, I settled for this. I now want to become the person, to inhabit the environment, that does the thing.

This book didn't exactly change my life, but it made me feel better about what I was already doing. (Before, I'd been calling it slingshot akrasia.)

**Structured procrastination is that staple from stand-up comedy where the best way to get your**

All of my reviews, all of my essays were written in the glow and shadow of other things I should've been doing.

All procrastinators put off things they have to do. Structured procrastination is the art of making this bad trait work for you. The key idea is that procrastinating does not mean doing absolutely nothing. Procrastinators seldom do absolutely nothing; they do marginally useful things, such as gardening or sharpening pencils or making a diagram of how they will reorganize their files when they find the time. Why does the procrastinator do these things? Because accomplishing these tasks is a way of not doing something more important. If all the procrastinator had left to do was to sharpen some pencils, no force on earth could get him to do it. However, the procrastinator can be motivated to do difficult, timely, and important tasks, as long as these tasks are a way of not doing something more important... Doing those tasks becomes a way of not doing the things higher on the list. With this sort of appropriate task structure, you can become a useful citizen. Indeed, the procrastinator can even acquire, as I have, a reputation for getting a lot done. Procrastinators often follow exactly the wrong tack. They try to minimize their commitments, assuming that if they have only a few things to do, they will quit procrastinating and get them done. But this approach ignores

the basic nature of the procrastinator and destroys his most important source of motivation. The few tasks on his list will be, by definition, the most important. And the only way to avoid doing them will be to do nothing. This is the way to become a couch potato, not an effective human being... The second step in the art of structured procrastination is to pick the right sorts of projects for the top of the list. The ideal projects have two characteristics – they seem to have clear deadlines (but really don't), and they seem awfully important (but really aren't). Luckily, life abounds with such tasks. At universities, the vast majority of tasks fall into those two categories, and I'm sure the same is true for most other institutions... At this point, the observant reader may feel that structured procrastination requires a certain amount of self-deception, since one is, in effect, constantly perpetrating a pyramid scheme on oneself. Exactly... what could be more noble than using one character flaw to offset the effects of another?

– Work and study pressurise my life. They give me a structure to defy, a gravity assist. I am happiest laden with obligations, when the set of tasks that is my life flies just out of control. I think the mechanism is this: 1. I require a steady stream of variety. 2. Having a job makes my days closely resemble each other. 3. Intolerable resentment. I am forced to produce creative sparks to satisfy my basic drive. SP is related to how great I feel when I don't have to go to a party, to my sadly efficient approach to grades, to how giving work to a busy person is a good way of getting it done quicker, i.e. an implausible linear increase of output with increasing things to do. I read more fiction when doing a stats degree and learn more stats when in work. — Antecedents of Perry and me. Fernando Pessoa:

I often wonder what kind of person I would be if I had been protected from the cold wind of fate by the screen of wealth... to reach the tawdry heights of being a good assistant book-keeper in a job that is about as demanding as an afternoon nap and offers a salary that gives me just enough to live on. I know that, had that past existed, I would not now be capable of writing these pages, which, though few, I would undoubtedly have only day-dreamed, given more comfortable circumstances. For banality is a form of intelligence, and reality, especially if it is brutish and rough, forms a natural complement to the soul. Much of what I feel and think I owe to my work as a book-keeper since the former exists as a negation of and flight from the latter.

Nietzsche:

the struggle against the ecclesiastical oppression of millenniums of Christianity... produced in Europe a magnificent tension of soul, such as had not existed anywhere previously; with such a tensely-strained bow one can now aim at the furthest goals... we have it still, all the distress of spirit and all the tension of its bow! And perhaps also the arrow, the duty, and, who knows? The goal to aim at...

Geoff Dyer:

The best circumstance for writing, I realized... were those in which the world was constantly knocking at your door; in such circumstances, the work you were engaged in generated a kind of pressure, a force to keep the world at bay. Whereas here, on Alonissos, there was nothing to keep at bay, there was no incentive to generate any pressure within the work, and so the surrounding emptiness invaded and dissipated, overwhelmed you with inertia. All you could do was look at the sea and the sky and after a couple of days you could scarcely be bothered to do that.

Zach Weiner:

[After months of working only on my main goal] I took on a job doing closed captioning because I found it [made for] an easier time writing. Just something about talking to people and watching weird media made the writing a lot easier. My new theory of self was that you can't write well unless you have a little strife in your life. I worked at the closed captioning job for 4-6 months and by then I was making enough money on the site to responsibly quit my job. The problem was I didn't want to quit my job and have readership fall off because I couldn't write, so my crazy idea was to go back to school. I thought, it'd be this weird environment, with younger people, and that would be good...

— Is this platitudinous? It is possible that the grand narration above is delusional, and that the only actual content here is “A lot of people work better under pressure”. Don’t think so though.

YMMV. 5/5 if you don’t do this already

```
<h3>To medawar</h3>
<div>
  <blockquote>
    A danger sign... is the tendency to regard the happiest moments of your life as tho
  </blockquote>
  <center>- Peter Medawar</center>
  <br><br>

  Another way I am ruled by demons: I have rarely had a shock of recognition stronger than
  You go along with it, screw up your courage, walk down the road - and find the Tube line
  The other says: "Joy of joys! - life pouring in rivulets down my chin! I am master of t

  It doesn't work if <i>you</i> cancel. The game is not that easy. Instead, you must be a
</div>

<h3>Why listen to me on this topic?</h3>
<div>
  <i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect
  <ol>
    <li>immersion in the field and/or good priors for what makes for an extraordinar
    <li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes
    <li>incredible amounts of argument-checking, which doesn't need domain knowledge
  </ol>
</div>
```

</ol>

I always try to do (3) but surely often fail.</i> <br><br><br>

In this case, anecdotal reasons: I am a natural practitioner of the method and think it  
</div>

*Cross-posted from Goodreads.*

# Reversals in psychology

Gavin

2020-01-26

{% include psy/links.md %}

Now a crowdsourced project elsewhere. Seeking volunteers!

A medical reversal is when an existing treatment is found to actually be useless or harmful. Psychology has in recent years been racking up reversals: in fact only 40-65% of its classic social results were replicated, in the weakest sense of finding ‘significant’ results in the same direction. (Even in those that replicated, the average effect found was half the originally reported effect.) Such errors are far less costly to society than medical errors, but it’s still pollution, so here’s the cleanup.<sup>1</sup>

Psychology is not alone: medicine, cancer biology, and economics all have many irreproducible results. It’d be wrong to write off psychology: we know about most of the problems here because of psychologists, and its subfields differ a lot by replication rate and effect-size shrinkage.

One reason psychology reversals are so prominent is that it’s an unusually ‘open’ field in terms of code and data sharing. A less scientific field would never have caught its own bullshit.

The following are empirical findings about empirical findings; they’re all open to re-reversal. Also it’s not that “we know these claims are false”: failed replications (or proofs of fraud) usually just challenge the evidence for a hypothesis, rather than affirm the opposite hypothesis. I’ve tried to ban myself from saying “successful” or “failed” replication, and to report the best-guess effect size rather than play the bad old Yes/No science game.<sup>2</sup>

Figures correct as of March 2020; I will put some effort into keeping this current, but not that much. Code for converting means to Cohen’s  $d$  and Hedge’s  $g$  here.

---

## Social psychology

{% include psy/social.md %}

## **Positive psychology**

{% include psy/positive.md %}

## **Cognitive psychology**

{% include psy/cognitive.md %}

## **Developmental psychology**

{% include psy/developmental.md %}

## **Personality psychology**

- Pretty good? One lab's systematic replications found that effect sizes shrank by 20% though. See the comments for someone with a fundamental critique.
- Anything by Hans Eysenck should be considered suspect, but in particular these 26 'unsafe' papers (including the one which says that reading prevents cancer).

## **Behavioural science**

- The effect of "nudges" (clever design of defaults) may be exaggerated in general. One big review found average effects were six times smaller than billed. (Not saying there are no big effects.)
- Here are a few cautionary pieces on whether, aside from the pure question of reproducibility, behavioural science is ready to steer policy.
- Moving the signature box to the top of forms does not decrease dishonest reporting in the rest of the form.

## **Marketing**

- Brian Wansink accidentally admitted gross malpractice; fatal errors were found in 50 of his lab's papers. These include flashy results about increased portion size massively reducing satiety.

## **Neuroscience**

{% include psy/neuro.md %}

## **Psychiatry**

- At most extremely weak evidence that psychiatric hospitals (of the 1970s) could not detect sane patients in the absence of deception.

## Parapsychology

- No good evidence for precognition, undergraduates improving memory test performance by studying after the test. This one is fun because Bem's statistical methods were "impeccable" in the sense that they were what everyone else was using. He is Patient Zero in the replication crisis, and has done us all a great service. (Heavily reliant on a flat / frequentist prior; evidence of optional stopping; forking paths analysis.)

Stats

```
<ul>
    <li><span class="b">Original paper</span>: '<a href="{{bem0}}>Feeling the future
        <br>(&#126;1000 citations, but mostly not laudatory).
    </li><br>
    <li><span class="b">Critiques</span>: <a href="{{bem}}>Ritchie 2012</a>, n=150.
        <a href="{{slate}}>Gelman 2013</a>; <a href="{{schimm}}>Schimmack 2018</a>,
        <br>(total citations: 200)
    </li><br>
    <li><span class="b">Original effect size</span>: Various, mean d=0.22. For experi
        </li><br>
        <li><span class="b">Replication effect size</span>: Correlation between r= minus
    </ul>
```

## Evolutionary psychology

```
{% include psy/evo.md %}
```

## Psychophysiology

- At most very weak evidence that sympathetic nervous system activity predicts political ideology in a simple fashion. In particular, subjects' skin conductance reaction to threatening or disgusting visual prompts - a noisy and questionable measure.

Stats

```
<ul>
```

Original paper: Oxley et al, n=46 ( citations). p=0.05 on a falsely binarised measure of ideology.

Critiques: Six replications so far (Knoll et al; 3 from Bakker et al) , five negative as in nonsignificant, one forking ("holds in US but not Denmark") (total citations: )

Original effect size: [ ], n=

Replication effect size: [ ], n=

## Behavioural genetics

```
{% assign intel = "https://pubmed.ncbi.nlm.nih.gov/23012269/" %} {% assign
schizo = "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4414705/pdf/nihms653267.pdf"
%} {% assign allofit = "https://www.eurekaselect.com/59624/article" %}
```

- No good evidence that 5-HTTLPR is strongly linked to depression, insomnia, PTSD, anxiety, and more. See also COMT and APOE for intelligence, BDNF for schizophrenia, 5-HT2a for everything...
- Be very suspicious of any such “candidate gene” finding (post-hoc data mining showing large >1% contributions from a single allele). 0/18 replications in candidate genes for depression. 73% of candidates failed to replicate in psychiatry in general. One big journal won’t publish them anymore without several accompanying replications. A huge GWAS, n=1 million: “*We find no evidence of enrichment for genes previously hypothesized to relate to risk tolerance.*”

[What I propose] is not a reform of significance testing as currently practiced in soft-psych. We are making a more heretical point... We are attacking the whole tradition of null-hypothesis refutation as a way of appraising theories... Most psychology using conventional H\_0 refutation in appraising the weak theories of soft psychology... [is] living in a fantasy world of “testing” weak theories by feeble methods.

– Paul Meehl (1990)

What now? When the next flashy WEIRD paper out of a world-class university arrives, will we swallow it?

Andrew Gelman and others suggest deflating *all* single-study effect sizes you encounter in the social sciences, without waiting for the subsequent shrinkage from publication bias, measurement error, data-analytic degrees of freedom, and so on. There is no uniform factor, but it seems sensible to divide novel effect sizes by a number between 2 and 100 (depending on its sample size, method, measurement noise, *maybe* its p-value if it’s really tiny)...

{% include psy/caveats.md %} {% include psy/foots.md %}

## See also

- A review of 2500 social science papers, showing the lack of correlation between citations and replicability, between journal status and replicability, and the apparent lack of improvement since 2009.
- Discussion on Everything Hertz, Hacker News, Andrew Gelman, some star data thugs comment.

*Thanks to Andrew Gelman, Stuart Ritchie, Anne Scheel, Daniël Lakens, Gwern Branwen, and Nick Brown for pointers to effectively all of these.*

*All honour to the hundreds of data thug / methodological terrorist psychologists I’ve cited, who in the last decade began the hard work of cleaning up their field.*

# ‘Peter Watts is an Angry Sentient Tumor’ (2019)

Gavin

2020-02-08

```
{% assign hogans = “https://en.wikipedia.org/wiki/Krista_and_Tatiana_Hogan”
%} {% assign blog = “https://www.rifters.com/crawl/?p=8812” %} {% assign
stross = “http://www.antipope.org/charlie/blog-static/” %} {% assign
bem = “https://replicationindex.com/2018/01/05/bem-retraction/” %} {% assign
hydro = “https://www.gwern.net/Hydrocephalus” %} {% assign crisis
= “https://www.americascientist.org/article/the-statistical-crisis-in-science” %}
{% assign lap = “https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00117/full”
%} {% assign bay = “https://rationalwiki.org/wiki/Extraordinary_claims_require_extraordinary_evidence#Pro
%} {% assign hals = “https://docs.google.com/document/d/1qmHh-
cshTCMT8LX0Y5wSQm8FMBhaxhQ8OIeRLkXIF0” %} {% assign kel =
“https://www.vox.com/future-perfect/2019/6/13/18660548/climate-change-
human-civilization-existential-risk” %} {% assign ozy = “https://forum.effectivealtruism.org/posts/eJPjSZKyT4t
change-is-in-general-not-an-existential-risk” %} {% assign sand = “https://www.newsweek.com/will-
climate-change-wipe-out-humanity-opinion-1440384” %} {% assign ssc
= “https://slatestarcodex.com/2014/08/14/beware-isolated-demands-for-
rigor” %} {% assign sdg = “https://sdg-tracker.org” %} {% assign
story = “https://www.gwern.net/Story-Of-Your-Life” %} {% assign op
= “https://www.theguardian.com/tv-and-radio/2013/aug/09/oprah-winfrey-
swiss-apology-racist-treatment” %} {% assign morg = “https://www.goodreads.com/review/show/2297137492”
%} {% assign conf = “https://slatestarcodex.com/2018/01/24/conflict-vs-
mistake” %} {% assign beat = “https://www.rifters.com/crawl/?p=932” %} {% assign
necro = “https://www.rifters.com/crawl/?p=1831” %}
```

Eleven years after the birth of the most neurologically remarkable, philosophically mind-blowing, transhumanistically-relevant being on the planet, we have nothing but pop-sci puff pieces and squishy documentaries to show for it. Are we really supposed to believe that in over a decade no one has done the studies, collected the data, gained any insights about literal brain-to-brain communication, beyond these fuzzy generalities? I for one don’t buy that for a second. These neuroscientists smiling at us from the screen—Douglas Cochrane, Juliette Hukin—they know what they’ve got. Maybe they’ve discovered something so horrific about the nature of Humanity that they’re afraid to reveal it, for fear of outrage and widespread panic. That would be cool.

Selected blogposts from a thoughtful doomer. Name a hot button, anything,

and Watts will elevate it to the scariest thing in the world: internet surveillance, zoonotic viruses, climate change, Trump, the security detail around the G8.

There's much to like: his bloody-minded sympathy, Left nihilism, boundless sensawunda, viscera instead of prose - and but deep unreliability when he gets on a subject besides marine biology. He is vulnerable to anything cool or fucked up. I worry if I find myself agreeing with him, since he so often misleads himself.

If I am indeed fated to sink into this pit of surveillance capitalism with the rest of you, I'd just as soon limit my fantasies about eating the rich to a venue that doesn't shut you down the moment some community-standards algo thinks it sees an exposed nipple in a jpeg.

Everything he does is excessive. Of course, this makes for good aesthetics and bad epistemics.

Like Charlie Stross, Watts reads horrifying things into the news, informed by the toxic half of history but also by a nebulous paranoia which leads them astray. (Representative sample from Stross: “[media incentive] has been weaponized, in conjunction with data mining of the piles of personal information social networks try to get us to disclose (in the pursuit of advertising bucks), to deliver toxic propaganda straight into the eyeballs of the most vulnerable — with consequences that are threaten to undermine the legitimacy of democratic governmance on a global scale.” Watts:

Bureaucratic and political organisms are like any other kind; they exist primarily to perpetuate themselves at the expense of other systems. You cannot convince such an organism to act against its own short-term interests... It's not really news, but we seem to be living in a soft dictatorship. The only choices we're allowed to make are those which make no real difference... On a purely selfish level I'm happier than I've ever been in my life, happier than I deserve. Of course it won't last. I do not expect to die peacefully, and I do not expect to die in any jurisdiction with a stable infrastructure. At least I don't have to worry about the world I'm leaving behind for my children; I got sterilized in 1991.

)

The two biggest fumbles here are his posts on Daryl Bem and high-functioning hydrocephalic people. It is no shame to fall for either: these are highly respectable academic errors (not hoaxes), and Bem's methods were exactly as valid as the average psychology paper of the early C21st. Watts' mistake isn't to insist that ESP is real, but to leap to the defence of the weird *just because it is weird*, to the point where he rejects Hume's maxim (“Laplace's principle”), a basic incontrovertible theorem of Bayesian inference.

[Bem's] results, whatever you thought of them, were at least as solid as those used to justify the release of new drugs to the consumer market. I liked that. It set things in perspective, although in hindsight, it probably said more about the abysmal state of Pharma regulation... I'm perfectly copacetic with the premise that psychology is broken. But if the field is really in such disrepair, why is it

that none of those myriad less-rigorous papers acted as a wake-up call? Why snooze through so many decades of hack analysis only to pick on a paper which, by your own admission, is better than most?

The question, here in the second decade of the 21st Century, is: what constitutes an “extraordinary claim”? A hundred years ago it would have been extraordinary to claim that a cat could be simultaneously dead and alive; fifty years ago it would have been extraordinary to claim that life existed above the boiling point of water, kilometers deep in the earth’s crust. Twenty years ago it was extraordinary to suggest that the universe was not only expanding but that the rate of expansion was accelerating. Today, physics concedes the theoretical possibility of time travel

Another big miss is his emphasis on adaptive sociopathy as the cause of our problems, rather than say lack of global coordination power. He is also completely off the deep end on climate change as existential risk, sneering at anyone who disagrees, no matter how well-informed. (He’s far from alone in that.)

there’s no denying that pretty much every problem in the biosphere hails from a common cause. Climate change, pollution, habitat loss, the emptying of biodiversity from land sea and air, an extinction rate unparalleled since the last asteroid and the transformation of our homeworld into a planet of weeds—all our fault, of course. There are simply too many of us. Over seven billion already, and we still can’t keep it in our pants.

Notice the pattern: faced with an apparent dilemma, he happily chuck the strongest, most basic principles to maintain his paranoia (the principles “extraordinary claims require extraordinary evidence” or here “it is good for people to have children if they want, good lives have worth”).

This bias would be entirely fine if he only admitted error later, about his predicted Trump race riots for instance.

The real danger isn’t so much Trump himself, but the fact that his victory has unleashed and empowered an army of bigoted assholes down at street level. That’s what’s gonna do the most brutal damage.

Most of these posts are entertaining but betray one-way critical thinking: for some reason he can barely see the other half of the world, that we are winning in all kinds of ways.

Lots of learned and fun film reviews: I relaxed, since criticism need have no truth-value. He likes ‘Arrival’ more than ‘Story of Your Life’, which fits: the film is bombastic, paranoid, politicised, unsubtle.

When you can buy the whole damn store and the street it sits on with pocket change; when you can buy the home of the asshole who just disrespected you and have it bulldozed; when you can use your influence to get that person fired in the blink of an eye and turn her social media life into a living hell—the fact that you don’t do any of those things does not mean that you’ve been oppressed.

It means you've been merciful to someone you could just as easily squash like a bug... Marvel's mutants are something like that. We're dealing, after all, with people who can summon storm systems with their minds and melt steel with their eyes. Xavier can not only read any mind on the planet, he can freeze time, for fucksake. These have got to be the worst case-studies in oppression you could imagine.

it still seems a bit knee-jerky to complain about depictions of objectification in a movie explicitly designed to explore the ramifications of objectification. (You could always fall back on Foz Meadows' rejoinder that "Depiction isn't endorsement, but it is perpetuation", so long as you're the kind of person who's willing to believe that Schindler's List perpetuates anti-Semitism and The Handmaid's Tale perpetuates misogyny.)

Watts reacts with caution and indignation to any police presence, even a compassionate visit to the homeless man sleeping in his garden. It would be crude to explain away Watts' style and worldview by reference to his unusually bad luck: his flesh-eating disease, his senseless beatdown and prosecution by border cops, his publishing travails, his scientific and romantic flops.

I'm probably only so down on him because I got so excited by Blindsight and its promise of actual science fiction by an actual scientist. He is certainly well above-average rigour for a political blogger, and well above-average imagination for anyone.

```
<h3>I was promised fictional luxury space communism</h3>
<div>
Peter Watts; Charlie Stross; <a href="{morg}">Richard Morgan</a>; William Gibson; NK Jemis
A glib answer is that they (cynically: their audience) feel they have lost the present (to r
I hear Chomskyan echoes in the blogs of the above, so maybe his analytical pessimism is also
Does solarpunk have anything worth reading yet?
</div>

<h3>Why listen to me on this topic?</h3>
<div>
    <i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect
    <ol>
        <li>immersion in the field and/or good priors for what makes for an extraordinar
        <li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes
        <li>incredible amounts of argument-checking, which doesn't need domain knowledge
    </ol>
    I always try to do (3) but surely often fail.</i> <br><br><br>
    In this case, I am probably about as trustworthy as Watts. Though I am only half a scien
</div>
```

# ‘Starting Strength’ by Rippetoe

Gavin

2020-02-22

```
{% assign mac = "https://sandymaguire.me/blog/no-coffee/" %} {% assign factory = "https://www.lesswrong.com/posts/8rdoea3g6QGhWQtmx/existential-angst-factory" %}
```

... a life is like iron. If you make good use of it, it wears out; if you don't, rust destroys it. So too we see men worn out by toil; but sluggishness and torpor would hurt them more.

– Cato the Elder

The first paragraph of this fitness book has stronger writing than you'd ever expect:

Physical strength is the most important thing in life. This is true whether we want it to be or not... Whereas previously our physical strength determined how much food we ate and how warm and dry we stayed, it now merely determines how well we function in these new surroundings we have crafted for ourselves as our culture has accumulated. But we are still animals – our physical existence is, in the final analysis, the only one that actually matters. A weak man is not as happy as that same man would be if he were strong. This reality is offensive to some people who would like the intellectual or spiritual to take precedence. It is instructive to see what happens to these very people as their squat strength goes up.

It begins with a metaphysical salvo(!)

This message is repulsive, unjust, and almost exactly fits my experience. (Though he's being imprecise: better to say “the most important foundation”, that fitness is a key instrument rather than the highest terminus. Though even then it's not “most important”, since it neglects even larger nonintellectual effects on my philosophy of life: love, and grand moral scheming.)

There is lots of reasoning from first principles, which is satisfying and gives it an Athenian air, but which I can just barely evaluate. Luckily it is just so easy to check if he's right (for your case).

The force of gravity acting on the bar is always acting straight down in a vertical line. Therefore, the most efficient way to oppose this force is by acting on it

vertically as well. So not only is a straight line the shortest distance between two points, but a straight vertical line is also the most efficient bar path for a barbell moving through space in a gravitational framework.

Your bench press strength doesn't adapt to the total number of times you've been to the gym to bench or to your sincerest hope that it will get stronger. It adapts to the stress imposed on it by the work done with the barbell. Furthermore, it adapts to exactly the kind of stress imposed on it. If you do sets of 20, you get good at doing 20s. If you do heavy singles, you get better at doing those.

"good technique" in barbell training is easily and understandably defined as the ability of the lifter to keep the bar vertically aligned with the balance point.

Rippetoe is the source of the recent renaissance in cheap simple barbells (free weights, i.e. dozens of muscles recruited at once) over circuits of giant single-muscle machines. (In the West, anyway; large parts of the world, e.g. Russia, apparently never gave up their bars.) He tells a plausible mean story about the perverse economic incentives that led to the latter, 1980-2010.

There is too much detail here - he discusses variants of the movements and the debate over them. But what a trivial criticism that is! I think most people could skip two-thirds of the book, since there's detailed kinematics for each move, instructor tips and gym-building tips, but it's interesting throughout. You could get the key parts from the final Programme section, then the "what not to do" chapter closing sheets. Warm-up sets chapter was very useful.

if your schedule does not allow time for proper warm-up, it does not allow time for training at all... [The squat] should be carefully and thoroughly prepared with a couple of empty-bar sets, and then as many as five sets between those and the work sets.

There's an abrupt shift in tone, in the chapter on lifting for kids: he starts citing University press books and listing comparative numbers for his claims. So this is a crusade for him. It is unlikely that you'd learn form from this alone, even like reading it and applying it live with a mirror. It is unlikely that you could find a PT with this much physical knowledge or clarity. He's quite bitchy, which I like but you might not:

if you continually miss workouts, you are not actually training, and your obviously valuable time should be spent more productively elsewhere.

If you're not increasing your max weight lifted, you're not training, and so not following his programme. The obsession with increase is still not mine. Strength, yes, exertion yes, but constant expansion? I aim for 100kg squat, and expect to attain it this year. Not herniating weight, not kneecapping weight, not sclerotic weight: nice big weight. Maybe once I get there I will grow bored, will again be confounded by the power of concrete body on worldview, and have to start climbing again. He thinks everyone gets injured eventually. But is this under the permanent revolution programme?

[Since writing this I looked quite hard and couldn't find any evidence that injury rates increase with (slowly attained) weight, until you get to the crazy competition levels.]

Ambition is useful, greed is not. Most of human history and the science of economics demonstrate that the desire for more than is currently possessed drives improvement, both personally and for societies. But greed is an ugly thing when uncontrolled and untempered with wisdom, and it will result in your progress coming to an ass-grinding halt.

If you're a little fluffy around the belly, you have obviously already created the conditions necessary for growth. You'll usually start out stronger than the skinny guy, and because your body hasn't got the problems with growing that skinny guys do, strength gains can come more easily for you if you eat correctly.

I've been doing a derivative of this program since October last year: no trainer, lots of missed sessions, just the primary exercises, 1 hour and out, a scaled-up ordinary diet, and I still saw decent gains, +50kg onto my initial squat. Rippetoe claims this could be achieved in half the time with many gallons of milk and much more aggro, and I see no reason to doubt him.

## Philosophical aspects of lifting

What sort of philosophy one chooses depends... on what sort of man one is; for a philosophical system is not a dead piece of furniture that we can reject or accept as we wish; it is rather a thing animated by the soul of the person who holds it.

— Fichte

Gradually it becomes clear to me what every great philosophy really was – namely the confession of its originator, and a sort of involuntary and unconscious autobiography...

— Nietzsche

- As above: The body helps determine the mind. You should be wary of your own philosophy, not just because of your local social conditioning, but also because of your diet, your habits, your daily kindness, and your bench. The lifter is Sisyphus, happy.
- Weights are a strong psychological intervention, perhaps the third-strongest I have found, after love and higher purpose. It's comical, how much of my deep teenage unhappiness, and sincere existentialism, was grounded in concrete fixable problems, and how little I understood that they were both fixable and not in fact intellectually grounded. Not knowing how to talk to girls, not exercising, not actively helping people: these produced my philosophy. Now *that's* absurdism!

- No excuses, no wiggle room, no ambiguity: lifting a lot without injuring yourself is a brute fact, unbiased. Rippetoe: “cause and effect cannot be argued with or circumvented by your wishes and desires.”
- ‘He’s a growing loun!’ my granny would say, justifying my early gluttony. Well, twenty years later here I am again, a growing boy. Artificial growth, body neoteny. What does a sense of increase, of coming potential, do to you?
- Waiting until soreness subsides before doing the next workout is a good way to guarantee that soreness will be produced every time, since you’ll never get adapted to sufficient workload frequency to stop getting sore.
- There are so many ways to do it wrong. (Only some of those wrong ways break you - the others just slow you down or confuse your body.) Rippetoe focusses on five movements, out of however many thousand physiologically possible ones. These are picked for excellent reasons, tested over decades.
- Exercise is the thing we must do to replicate the conditions under which our physiology was – and still is – adapted, the conditions under which we are physically normal. In other words, exercise is substitute caveman activity, the thing we need to make our bodies, and in fact our minds, normal in the 21st century.
- Psychologically, 20 [rep max] work is very hard, due to the pain, and lifters who are good at it develop the ability to displace themselves from the situation during the set. Or they just get very tough.
- I live in my head. But the hip drive out of a deep squat is such a strong strange confluence of forces, vaguely under my control but more accurately an explosion I light the fuse on, that I am driven to notice and appreciate neuromuscular marvels.

<h3>Why listen to Rippetoe on this topic?</h3>

<div>

Decades of personal experience plus strong amateur theory plus distilled folk wisdom, in

<blockquote>

*<i>Most sources within the heavy-training community agree that a good starting place*

</blockquote>

</div>

<h3>Why listen to me on this topic?</h3>

<div>

*<i>Nonfiction book reviews by nonspecialists are hazardous. It is just not easy to detect*

<ol>

*<li>immersion in the field and/or good priors for what makes for an extraordinary*

*<li>incredible amounts of fact-checking gruntwork, at least 5x the time it takes*

*<li>incredible amounts of logic-checking, which doesn't need domain knowledge.</i>*

</ol>

*I always try to do (3) but surely often fail.</i> <br><br><br>*

I've followed <a href="https://stronglifts.com/">Rippetoe's programme</a> inconsistently  
</div>

# AI ethics for present & future

Gavin

2020-05-30

{% include ai-ethics/links.md %}

Professional physicists who investigate the first three minutes or the first microsecond no longer need to feel shy when they talk about their work. But the end of the universe is another matter... the striking thing about these papers is that they are written in an apologetic or jocular style, as if the authors were begging us not to take them seriously. The study of the remote future still seems to be as disreputable today as the study of the remote past was thirty years ago.

— Freeman Dyson (1979)

AI ethics (a family of fields including ‘Fairness, Accountability, & Transparency in ML’, ‘robot ethics’, ‘machine ethics’, and ‘AI law’) is awash with money and attention following the last decade’s enormous progress in AI systems’ performance. Just at my own university, Bristol, I count 5 researchers who have begun on this topic, on aspects like the ethics of self-driving vehicles in dangerous situations and the ethics of emotionally responsive robots, including carers, pets, and lovers.

In addition, parallel work focusses on a technology which does not yet exist: artificial general intelligence (AGI), that is, a system which could do anything a human can do, and maybe more. The issues around such a technology are quite different from the short-term issues with present pattern-matching AI systems. If they were realised, such systems could transform society through the automation of almost all labour, including the scientific and engineering labour which is so often the limiting factor in economic progress, and could even carry a risk of accidental human extinction (‘existential risk’).

Call this trend ‘AGI safety’. It has been increasing in prominence, and some of the most respected CS researchers now take the idea seriously, including Stuart Russell, the author of the most prominent textbook in AI.

If the two trends were marked only by a division of labour, there would be no problem: both scales are important, and both merit careful research. However,

there appears to be a degree of animosity and very little co-operation between the two clusters of research.

People talk past each other here. Elon Musk, Lord Martin Rees, and other famous figures have weighed in on existential risk from advanced AI: as a result, popular discussion of the issue focusses on rebutting informal versions of the longtermist argument. If you've encountered this debate, it's probably only the sensational form, or that plus the trivial counter-sensational pieces.

In fact, a growing minority of technical AI experts are openly concerned with the long-term impact. But when AI ethicists do acknowledge AGI safety, it is only by reference to figures outwith technical AI: the industrialists Elon Musk and Bill Gates, the philosopher Nick Bostrom - if we're lucky and the interlocutor isn't instead a static image of the T-800 robot from *Terminator*. The foil is never Turing Award winners Yoshua Bengio or Judea Pearl, Stuart Russell, or the prominent deep learning researcher Ilya Sutskever.

This sort of division is nothing new; as Dyson notes above, the tension between verification and speculation, direct evidence and extrapolation, short-run and long-run importance plays out in many fields. Academia is in general content to stick to the facts and the present context, and so to leave futurism to popular writers beyond the pale.

(There is actually a small literature on this very question, mostly taxonomies of disagreement and pleas for co-operation: Cave, Stix & Maas, Prunkl & Whittlestone, Krakovna.)

I think part of this is down to failures of communication, and part down to academia's natural, often helpful pre-emptive dismissal of weird ideas. Let's try and patch the first one.

### Why on earth might we worry about AGI?

It seems odd for scientists to not only speculate, but also to act decisively about speculative things - for them to seem sure that some bizarre made-up risk will in fact crop up. The key to understanding this is expected value: if something would be extremely important if it happened, then you can place quite low probability on it and still have warrant to act on it.

Consider finding yourself in a minefield. If you are totally uncertain about whether there's a buried landmine right in front of you - not just "no reason to think so", you genuinely don't know - then you don't need direct evidence of it in order to worry and to not step forward.

The real argument is all about uncertainty: advanced AI systems could be built soon; they could be dangerous; making them safe could be very hard; and the combination of these probabilities is not low enough to ignore.

- When you survey technical AI experts, the average guess is a "10% chance of human-level AI (AGI)... in the 2020s or 2030s". This is weak evidence,

since technology forecasting is very hard; also these surveys are not random samples. But it seems like some evidence.

- We don't know what the risk of AGI being dangerous is, but we have a couple of analogous precedents: the human precedent for world domination, at least partly through relative intelligence; the human precedent for 'inner optimisers', unexpected shifts in the goals of learned systems. Evolution was optimising genetic fitness, but produced a system, us, which optimises a very different objective ("fun; wellbeing"); there's a common phenomenon of very stupid ML systems still developing "clever" unintended / hacky / dangerous behaviours.
- We don't know how hard alignment is, so we don't know how long it will take to solve. It may involve hard philosophical and mathematical questions.

One source of confusion is the idea that the systems would have to be malevolent, intentionally harmful, to be dangerous; Nick Bostrom's much-misunderstood 'paperclip maximiser' argument shows one way for this to be untrue: when your AI system is a *maximiser*, as for instance almost all present 'reinforcement learning' AI systems are, then bad effects can (and do) arise from even very minor mistakes in the setup.

Another involves equating intelligence with consciousness, missing that the AI notion of 'intelligence' is based on mere capacity for clever behaviour, and not any thorny philosophical questions of subjectivity or moral agency. This sidelines the very large (and for all I know valid) body of work from phenomenology criticising the very idea of machine consciousness.

It's not that the general idea is too extreme for the public, or world government. One form of existential risk is already a common topic of discussion and a core policy area: the possibility of extreme climate change. (But, while that risk is also marked by uncertainty, animosity, and distrust, this conflict is mostly outside academic boundaries.) And this follows the broad-based opposition to nuclear proliferation, perhaps the first mass movement against x-risk in history.

### **What might be wrong with taking the long-term view?**

Humans aren't very good at forecasting things more than a couple years ahead. To the extent that a given long-termist claim depends on precise timing, it isn't possible to pull off.

Weird ideas are usually wrong, and sadly often say something about the person's judgment in general.

Most gravely, if resources (funding, popular and political attention) are limited, then long-termism could be a distraction from current problems. Or worse, counterproductive, if we did short-term harm to promote an unsure longterm benefit.

### What might be wrong with taking the short-term view?

The long-run is much larger than the short-run, and could, all going well, contain many, many more people. On the assumption that future people matter at all, their well-being and chance to exist is the largest moral factor there can be; and even in the absence of this assumption, the premature end of the current generation would also be an extreme tragedy. Future people are the ultimate under-represented demographic: despite nice moves in a handful of countries, they have no representation.

Our choice of timeframe has intense practical consequences. From a short-term view, technology has many risks and only incremental benefits. But in the long run, it is our only hope of not dying out: at the very latest, because of the end of the Sun's lifespan.

The worry about counterproductive work from the section above applies equally to short-termism. It would be quite a coincidence if picking the thing which is most politically palatable, which improves matters in the short-run was also the thing that helped us most in the long-run. One example of a short-term gain which could have perverse long-term effect is the present trend towards national or (bloc) AI strategies in the pursuit of local (zero-sum) economic or military gain, which could easily lead to an AI 'arms race' in which safety falls by the wayside. That said, there are plenty of opportunities which seem robustly good on all views, like increasing the transparency of AI.

Sketch of a unified ethics of AI from Prunkl & Whittlestone.

Ultimately it's not a binary matter and there's no need for jostling. The figure above shows how to consider all of AI ethics on the same page, as a matter of degree, and encourages us to consider all the impacts.

Returning to the epigram from Dyson: there is hope. Since 1979, respectable work on the end of the universe has flourished. There remains a great deal of uncertainty, and so an array of live contradictory hypotheses - and quite right too. (For instance, Dyson's own early model was obsoleted by the discovery that the cosmic expansion is accelerating.) There is no antipathy between physicists studying the cosmic birthday and those working on the cosmic doomsday - and quite right too. Perhaps we can repeat the trick with AI safety. There are few places it is more important to avoid factional disdain and miscommunication.

<h3>What moral assumptions are you making?</h3>

<div>

Few. The general long-termist argument applies to a huge range of worldviews; it is quite

<!-- -->

On the object level, views which can ignore existential risk include: People with incredi

<!-- -->

On the second level, perhaps one assumption is that 'academia should do good'. (Not only

</div>

<h3>If you're so smart, why ain't you mainstream?</h3>

```

<div>
  It's a new idea and it has a bunch of baggage ("cached thoughts") from fiction.<br><br>
  Also, academia is conservative, in the sense that it pays almost all its attention to the
  <!-- -->
  Short-term bias resulting from naive empiricism and the need to maintain respectability
</div>
<!-- -->
<h3>Biblio</h3>
<div>
  <ul>
    <li>David Buchanan. '<a href="{{buch}}>No, the robots are not going to rise up and kill us</a>'</li>
    <li>Oren Etzioni. '<a href="{{etz}}>No, the experts don't think superintelligent AI is a threat</a>'</li>
    <li> Allan Dafoe and Stuart Russell. '<a href="{{russell}}>Yes, we are worried about the long-term risks</a>'</li>
    <li> Luke Muehlhauser. '<a href="{{luke}}>Reply to LeCun on AI safety</a>', 2016.</li>
    <li> Anders Sandberg. '<a href="{{sand}}>The five biggest threats to human existence</a>'</li>
    <li> Kelsey Piper. '<a href="{{elon}}>Why Elon Musk fears artificial intelligence</a>'</li>
    <li> Martin Rees. '<a href="{{tele}}>Astronomer Royal Martin Rees: How soon will robots exceed us?</a>'</li>
    <li> Stephen Cave and Sean Ó hEigearthaigh. '<a href="{{oheig}}>Bridging near-and long-term existential risk</a>'</li>
    <li> Charlotte Stix and Matthijs Maas. '<a href="{{maas}}>Crossing the gulf between 'near' and 'long-term' existential risk</a>'</li>
    <li> Carina Prunkl and Jess Whittlestone. '<a href="{{prunkl}}>Beyond near- and long-term existential risk</a>'</li>
    <li> Viktoriya Krakovna. '<a href="{{krak}}>Is there a trade-off between immediate and long-term existential risk?</a>'</li>
    <li> Gavin Leech. '<a href="{{xrisk}}>Existential risk as common cause</a>'</li>
    <li> Nick Bostrom. '<a href="{{bost}}>Existential risk prevention as global priority</a>'</li>
    <li> Freeman J. Dyson. '<a href="{{dyson}}>Time without end: Physics and biology in an infinite universe</a>'</li>
    <li> <a href="{{facct}}>ACM conference on fairness, accountability, and transparency</a>'</li>
    <li> Geoff Keeling, Katherine Evans, Sarah M Thornton, Giulio Mecacci, and Filippo Santambrogio. '<a href="{{keeling}}>Machine decisions and their consequences</a>'</li>
    <li> Teresa Scantamburlo, Andrew Charlesworth, and Nello Cristianini. '<a href="{{scantamburlo}}>Machine decisions and their consequences</a>'</li>
    <li> Christopher Burr and Nello Cristianini. '<a href="{{burr}}>Can machines read our minds? Minds and Machines</a>'</li>
    <li> Stuart Russell. '<a href="{{russell}}>Human Compatible: Artificial intelligence and the problem of control</a>'</li>
    <li> Kate Crawford and Ryan Calo. '<a href="{{crawford}}>There is a blind spot in ai research.Nature, 538(7625):433-435, 2016.</a>'</li>
    <li> David Krueger. '<a href="{{krueger}}>A list of good heuristics that the case for AI safety</a>'</li>
    <li> AI Impacts. '<a href="{{timeline}}>AI timeline surveys</a>', 2017.</li>
    <li> Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. '<a href="{{hubinger}}>Specification gaming examples in AI</a>'</li>
    <li> Viktoriya Krakovna. '<a href="{{krakovna}}>Specification gaming examples in AI</a>'</li>
    <li> Toby Ord. '<a href="{{ord}}>The Precipice: existential risk and the future of humanity, 2020.</a>'</li>
    <li> Gregory Lewis. '<a href="{{greg}}>The person-affecting value of existential risk</a>'</li>
    <li> Christian Borggreen. '<a href="{{fort}}>AI fortress europe?</a>', 2020.</li>
    <li> Yun Wang et al. '<a href="{{wang}}>Current observational constraints on cosmic doomsday</a>'</li>
  </ul>
</div>

```

## See also

- Regulation of AI as power grab

# Crossing the ocean of my ignorance

Gavin

2020-06-22

```
{% include phone_img.html %} {% include ignorance/links.md %} {% assign
chapman = "https://dspace.mit.edu/bitstream/handle/1721.1/41487/AI_WP_316.pdf"
%} {% assign soares = "http://mindingourway.com/stop-trying-to-try-and-try/"%
%} {% assign steinhardt = "https://cs.stanford.edu/~jsteinhardt/ResearchasaStochasticDecisionProcess.html"
%} {% assign holden = "http://theorangeduck.com/page/reproduce-their-results"
%}
```

... postpone reading Nietzsche for the time being; first study Aristotle for ten to fifteen years.

– Martin Heidegger 1

As a researcher, either you won't understand something and you will feel stupid and like a worm, or you will understand something and think it's too trivial and hence still feel like a worm.

– Simon Peyton Jones

I was much further out than you thought And not waving but drowning.

– Stevie Smith

What do you need, to do new things? Imagine you're a junior researcher; a scientist; a dry-lab scientist; a Machine Learning person. For good and bad reasons you want to publish in Deep Learning, a decade-old bandwagon which continues to steamroll your field. You're rolling in the deep. How do you get to work?

A natural answer is to start at the beginning: go read the underlying mathematics.

OK, say you go off and do that. You're not happy with your understanding: you can feel the aching gap in your knowledge of say linear algebra - that your looking at all those matrices *actively concealed* something important - but you figure it's enough for now.

It takes a month or six. Can you do new things now? No: you have to learn how to actually implement things. Brilliant people have built easy tools for you, so you learn one of those and reimplement some big papers. This is harder than

it sounds, and you actually don't manage to reproduce half of the results. You add 3d6 unease and self-doubt.

That takes a month or two. Can you do new things now? No: you need a good idea. Where do you get those? 'Related Work', I guess. You go read. Later, your mouldering bones are discovered at your desk, with 200 tabs open and the Colab Disconnected modal still burning on your screen.

So much of the foundations I do not understand, and it would take a lifetime to fully understand them (and by then I'd have forgotten the first bits). With such a burden, how does anyone do new work? Well, by not doing any such thing.

- you have to just start
- you'll learn it when you need it
- most research is not done alone
- most researchers don't remember the low-level stuff, and don't have to
- you don't have to focus on one thing
- forcing yourself to work on something has large costs

Even after we reject foundationalism, the practical problem remains: what to learn, and how? I've been trying to think new things for about 6 years, but only recently got any good at it. Here are some things that may have helped:

### Requisite attitudes

#### The Neurathian bootstrap

We are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction.

– Otto Neurath

Beginning at the beginning, craving absolute foundations, mostly leads to paralysis. Sometimes this is because it takes too long to reach the frontier from the foundation; sometimes it's because the foundation is missing or impossible.

*To live, you have to ignore things.* So bite off a chunk of reality and ignore the rest. Manuel Blum:

When working on a PhD, you must focus on a topic so narrow that you can understand it completely. It will seem at first that you're working on the proverbial needle, a tiny fragment of the world, a minute crystal, beautiful but in the scheme of things, microscopic. Work with it. And the more you work with it, the more you penetrate it, the more you will come to see that your work, your subject,

encompasses the world. In time, you will come to see the world in your grain of sand.

People don't talk enough about what they ignore. One exception: Andrew Gelman, one of the most influential statisticians alive, never bothered with measure theory, the deep generalisation / justification of probability theory.

The raft is our lack of fear at the lack of raft.

### Comparing down

The above isn't about impostor syndrome, except insofar as I delude myself that others are not ignorant. I take impostor syndrome to be the subjective feeling of being inadequate relative to those around you. I'm talking about the objective sense in which no one has anything more than a piece of the puzzle; and yet some of them still manage to do new things. (To get a sense of how rough the subjective and objective problem is, note that PhD study breaks a quarter or a half of the smart people who try.)

Anyway: I had a very distorted view of how much an average PhD actually knows. Just as an undergraduate degree only shows you once had a small degree of knowledge on one or two topics, so too getting postgraduate funding only means that you're not totally dense and callow. This is good news! Not-totally dense and callow people manage to do many of the coolest things.

### Unlearning education

Books should follow science; science should not follow books.

- Francis Bacon

I was lucky; by being born in the right time and right place, I got huge amounts of free education.

I was unlucky; an education was not what I actually needed; education trains you for the wrong task, in the wrong way. The ability to do research correlates with doing well on tests. But it is probably not well served by the current degree of optimising for tests, reading, and mere recall.

There are multiple mismatches: it focusses your attention on solved or toy things; it emphasises understanding old things rather than creating new things; it expects you to do your best, not to solve things; it mostly doesn't let you follow your curiosity; it mostly doesn't train you to handle the gross uncertainty of research. (Outside of mathematics, there is no marking scheme - not even peer review, not even awards at conferences. Maybe 10 years later you'll get some sense of whether you actually succeeded.)

*Question first, not books first.* Learning is best and most lasting when in the service of a goal you actually care about: not "better grades", not "impress distant superior", but "I want to build x". When it is part of you.

PhDs are still pretty artificial (they make you work ~alone, on one pre-specified topic which has to look sensible and follow an existing programme, with deadlines, and you're fed ideas), but at least their goal is not a total dead-end.

It's not easy to unlearn tutelage, but at some point in your first few actual projects you might manage it.

### Ideas are cheap

One useful piece of startup culture: “you have to ship”. It is *not* that your perfect idea is ruined by imperfect implementation: your idea is nothing until it exists; all implementations are an improvement over an idea. 2

### Mechanisms

The above is about fixing your head. This bit is about how the vastness of the ocean actually ends up not mattering:

### Abstraction as testimony

```
{% assign tcp = "https://www.joelonsoftware.com/2002/11/11/the-law-of-leaky-abstractions/" %}
```

Some abstractions actually allow you to ignore what's underneath; some boats don't sink that quick. I've been playing with the internals of Pytorch recently. How many people understand the Tensor class? A couple hundred probably, for say 100,000 users of it, and who knows? a billion downstream users. In fact, most good software is about shielding you from details: even the statement `a = 1` is pretty computationally complicated. The world couldn't work without the glory of testimony like this.

### Collaboration

Even once you've selected a level of abstraction, trusted the bulkheads to hold, you can still split the work further: laterally across co-authors who are good at different parts. This is division of labour again, one of the most powerful social forces.

The average paper now has about 5 authors. Some of this is down to a deflation of what it takes to count as an author, but the rest is good stuff. One (conceptually) simple solution to the replication crisis in social science would be to require a statistician to be on every project, at least in the experiment design.

### Momentum

Ideas generate ideas, success generates success.

In Spring, I worked on a coronavirus modelling project. In writing it I collected 15 major ideas that we didn't have time for, didn't have data for, which didn't

fit into the scope of that paper. One week after submitting it, a subset of that team wrote another paper on the methods used, including 3 or 4 completely novel ideas and tests and proofs. We could do this 3 or 4 more times without a hint of ‘salami slicing’, bad behaviour. If we could only sustain the energy.

### Slack

You waste years not being able to waste hours

— Amos Tversky

One of the perversities of academic life is the absence of slack: spare time for just playing around. I won’t go into this here (see here instead), but here’s a nice story. A young mathematician recently cracked a notorious problem as a side-project, no deadline, no particular expectation of success, almost an étude.

### you have to have a question

<https://acesunderglass.com/2020/06/09/where-to-start-research/>

### Teaching as learning

You think you know when you can learn, are more sure when you can write, even more when you can teach, but certain when you can program

— Alan Perlis

Consider the grad student in teaching mode... When the professor asks them questions, they’re Expected To Do Their Best; when the undergrad asks them questions, they’re just expected to answer. In the first case, they’re expected to try; in the second case, they’re assumed capable, an assumption that fades into the background.

— Nate Soares

The bureaucracies act as if you can only teach once you are a master. But I often feel that I don’t understand anything until I try to explain it to someone else - hence this blog. Yet another unforced error of ordinary education: you’re not allowed to learn through teaching until it’s over.

“You learn the prerequisite in the next course.” And I learn the prerequisite when I am allowed to teach the prerequisite.

---

### See also

- Laura Deming’s rage.
- Matt Might’s ways to fail
- I think this post obsoletes some of the above.

- Peyton Jones, ‘How to Write’
- Steinhardt, ‘Research as a Stochastic Decision Process’
- Abram Demski in the Field,
- Nerst, Decoupling
- Alexander, ‘Ars longa, vita brevis’
- Holden, ‘Why Can’t I Reproduce Their Results?’
- Soares, ‘Stop trying to try and try’

... in order to understand the Epic of Gilgamesh, you'll have to first comprehend the cave paintings and sculpture produced during the Upper Paleolithic. Without a full grasp of the cave paintings at Lascaux, you'll never be able to contextualize the oral tradition that produced Gilgamesh, leaving you without a full knowledge of the Septuagint, making your reading of Kierkegaard incomplete, making your reading of Heidegger & Derrida faulty. Of course, you'll need to learn Proto-Indo-European.

There are subtleties here, about data fumes, info hazards, idea inoculation, and poisoning the well. But unless you're working on very strange things these are unlikely to apply.

# Self-help, hard and soft

Gavin

2020-06-07

I get annoyed when I click a post and find that it's not a list of things to try, or a cool tool for reasoning or deciding - but instead high-context talk about inner concerns. But I don't endorse this annoyance, and I try to be less narrow.

The following may be helpful if you're like me, and want to understand what people are doing and so circumvent annoyance. (I think it beats "rationality" vs "post-rationality" because it drops the incipient in-group/out-group stuff and focusses on what people are actually doing.)

I began with a vague 1D idea: that there's "hard self-help" (actionable, crisp, objective) and soft self-help (contemplative, open, subjective). This is really not useful without using two dimensions, though:

External

Internal

Legible

Lifehacks

Techniques

Illegible

Attempts

Seeking

Examples:

External

Internal

Legible

supplements, exercise, Things I Use, Quantified Self, checklists, Anki, time management, Algos to Live By,

Models, CBT, bug lists, IDC, Focussing, this 2x2

Illegible

brainstorming, CoZe, Oblique Strategies, ‘Actually Try’, ‘Dive In’, going with your gut

Art, ‘stream entry’, Kōans, aphorism, Replacing Guilt

## Notes

- This abstraction clearly strains to cover its large, wild, continuous domain.
- The top axis covers a few things. It could be “Environment optimisation vs Inner alignment”, or “Action vs Contemplation”, or “Objective vs Subjective”.
- My original name for the left axis was “Algorithm vs Heuristic”.
- The idea of a “Rest Day” is a legible algorithm: 1) Make no appointments; 2) Do what you want. But each instance is different and highly illegible. “*Why am I doing this? Because.*” Outer loop is algo, the inner loop is heuristic.
- If it worked, biofeedback would be the exemplar of *legible-internal*.
- This is probably not the most significant difference between rationality and post-rationality: that might be “rationality realism vs anti-realism”.
- I have no idea where to put Circling. It’s right in the middle.

## Build your own typology

- *Particular vs General*: Is it targeting one bug or a raised ‘waterline’? Object vs Meta level? Incremental or revolution?
- *Objective vs Subjective*: Is the target your environment (including the body) or your mind?
- *Action vs Contemplation*: Does the idea involve doing things besides thinking or introspection?
- *Algorithmic vs Heuristic*: Is the idea a complete recipe or a general prompt?
- *Propositional vs Nonpropositional*: Is the target beliefs, skills, plans, or attitudes?
- *Cognitive vs Emotional*.
- *System 1 vs System 2*: Is the target implicit or explicit?
- *Scientific or Rational evidence*: What reason is there to expect it to work?
- *Some feedback or no feedback*: how easy is it to check if it works?

<h3>Free-associating some vaguely related distinctions</h3>

<div>

- Analytic vs Continental<br>
- Hard SF vs Soft SF<br>
- STEM vs Humanities<br>
- Exoteric vs Esoteric<br>
- System 2 vs System 1<br>
- <a href="<https://stefanfschubert.com/blog/2020/5/30/the-clarity-dimension-and-the-accessibility-dimension>">

# AI alignment & academia

Gavin

2020-07-29

{% include acais/links.md %}

A big reason for the EA focus on AI safety is its neglectedness:

... less than \$50 million per year is devoted to the field of AI safety or work specifically targeting global catastrophic biorisks.

80,000 Hours (2019)

... we estimate fewer than 100 people in the world are working on how to make AI safe.

80,000 Hours (2017)

Grand total: \$9.09m... [Footnote: this] doesn't include anyone generally working on verification/control, auditing, transparency, etc. for other reasons.

Seb Farquhar (2018)

... what we are doing is less than a pittance. You go to some random city... Along the highway you see all these huge buildings for companies... Maybe they are designing a new publicity campaign for a razor blade. You drive past hundreds of these... Any one of those has more resources than the total that humanity is spending on [AI safety].

Nick Bostrom (2016)

Numbers like these helped convince me that AI safety is the best thing to work on. I now think that these are underestimates, because of non-EA lines of research which weren't counted.

Use "EA safety" for the whole umbrella of work done at organisations like FHI, MIRI, DeepMind and OpenAI's safety teams, and by independent researchers. A lot of this - maybe a third - is conducted at universities; to avoid double counting I count it as EA and not academia.

The argument: 1. EA safety is small, even relative to a single academic subfield.  
2. There is overlap between capabilities and short-term safety work. 3. There

is overlap between short-term safety work and long-term safety work. 4. So AI safety is less neglected than the opening quotes imply. 5. There's a good chance that academia will do more safety over time, eventually dwarfing the contribution of EA.

## What's 'safety'?

EA safety is best read as about "AGI alignment": work on assuring that the actions of an extremely advanced system are sufficiently close to human-friendly goals.

EA focusses on AGI because weaker AI systems aren't thought to be directly tied to existential risk. However, Critch and Krueger note that "prepotent" - unstoppably advanced, but not necessarily human-level - AI could still pose x-risks. The potential for this latter type is key to the argument that short-term work is relevant to us, since the scaling curves for some systems seem to be holding up, and so might reach prepotence.

"ML safety" could mean making existing systems safe, or using existing systems as a proxy for aligning an AGI. The latter is sometimes called "mid-term safety".

In the following "AI safety" means anything which helps us solve the AGI control problem.

## De facto AI safety work

The line between safety work and capabilities work is sometimes blurred. A classic example is 'robustness': it is both a safety problem and a capabilities problem if your system can be reliably broken by noise. Transparency (increasing direct human access to the goals and properties of learned systems) is the most obvious case of work relevant to capabilities, short-term safety, and AGI alignment. As well as being a huge academic fad, it's a core mechanism in 6 out of the 11 live AGI alignment proposals recently summarised by Evan Hubinger.

More controversial is whether there's significant overlap between short-term safety and AGI alignment. All we need for now is: The mid-term safety hypothesis (weak form): at least some work on current systems will transfer to AGI alignment. Some researchers who seem to put a lot of stock in this view: Shah, Christiano, Krakovna, Olsson, Olah, Steinhardt, Amodei, Krueger. (Note that I haven't polled them; this is guessed from public statements and revealed preferences.)

Here are some alignment-relevant research areas dominated by non-EAs. I won't explain these: I use the incredibly detailed taxonomy (and 30 literature reviews) of Critch and Krueger (2020). Look there, and at related agendas for explanations and bibliographies.

- Transparency
- Robustness
- Interactive AI

- Calibration
- Formal verification
- Preference learning
- Modelling human cognition
- Safe handovers (AKA corrigibility)
- Assured Autonomy
- Open source game theory
- Multi-agent coordination
- Emergent communication
- Safe RL
- (Parts of) algorithmic fairness

These are narrowly drawn from ML, robotics, and game theory: this is just a sample of relevant work! Work in social science, psychology, moral uncertainty, or decision theory could be just as relevant as the above direct technical work; Richard Ngo lists many questions for non-AI people here.

Work in these fields could help directly, if the eventual AGI paradigm is not too dissimilar from the current one (that is, if the weak mid-term hypothesis holds). But there are also indirect benefits: if they help us to use AIs to align AGI; if they help to build the field; if they help convince people that there really is an AGI control problem (for instance, Victoria Krakovna's specification gaming list has been helpful to me in interacting with sceptical specialists). These imply another view under which much academic work has alignment value:

*The mid-term safety hypothesis (very weak form):* at least some work on current systems will probably help with AGI alignment in some way, not limited to direct technical transfer.

A natural objection is that most of the above areas don't address the AGI case: they're not even trying to solve our problem. I discuss this and other discounts below.

## Current levels of safety-related work

### How large is EA Safety?

Some overlapping lists:

- # people with posts on the Alignment Forum since late 2018: 94. To my knowledge, 37 of these are full-time.
- 80k AI Safety Google Group: 400, almost entirely junior people.
- Larks' great 2019 roundup contained ~110 AI researchers (who published that year), most of whom could be described as EA or adjacent.
- Issa Rice's AI Watch: "778" (raw count, but there's lots of false positives for general x-risk people and inactive people. Last big update 2018).

In the top-down model I start with all EAs and then filter them by interest in AI risk, direct work, and % of time working on safety. (EA safety has a lot of

hobbyists, for instance me.) The bottom-up model attempts a headcount.

### How large is non-EA Safety?

A rough point estimate gives 84k or 103k AI academics, with caveats summarised in the Guesstimate notes. Then define a (very rough) relevance filter:

$$\begin{aligned} \text{CS} &= \% \text{ of capabilities work that overlaps with short-term safety} \\ \text{SL} &= \% \text{ of short-term safety that overlaps with long-term safety} \end{aligned}$$

Then, we could decompose the safety-relevant part of academic AI as:

$$\text{SR} = (\% \text{ of AI work on capabilities} * \text{CS} * \text{SL}) + (\% \text{ of AI work on short-term safety} * \text{SL})$$

None of those parameters is obvious, but I make an attempt in the model (bottom-left corner).

Then the non-EA safety size is simply the field size \* SR.

This just counts academia, and just technical AI within that. It's harder to estimate the amount of industrial effort, but the AI Index report suggests that commercial AI research is about 10% as large as academic research (by number of papers, not impact). But we don't need this if we're just arguing that the non-EA lower bound is large.

### What's a good discount factor for de facto safety work?

In EA safety, it's common to be cynical about academia and empirical AI safety. There's something to it: the amount of paperwork and communication overhead is notorious; there are perverse incentives around publishing tempo, short-termism, and conformity; it is very common to emphasise only the positive effects of your work; and, as the GPT-2 story shows, there is a strong dogma about automatic disclosure of all work. Also, insofar as AI safety is 'pre-paradigmatic', you might not expect normal science to make much headway. (But note that several agent-foundation-style models are from academia - see 'A cursory check' below.)

This is only half of the ledger. One of the big advantages of academic work is the much better distribution of senior researchers: EA Safety seems bottlenecked on people able to guide and train juniors. Another factor is increased influence: the average academic has serious opportunities to affect policy, hundreds of students, and the general attitude of their field toward alignment, including non-academic work on alignment. Lastly, you get access to government-scale funding. I ignore these positives in the following.

## Model

Here's a top-down model arguing that technical AI academics could have the same order of effect as EA, even under a heavy impact discount, even when ignoring other fields and the useful features of academia. Here's an (incomplete)

bottom-up model to check if it's roughly sensible. As you can see from the variance, the output means are not to be trusted.

A “confidence” interval

Again, the model is conservative: I don't count the most prominent safety-relevant academic institutions (FHI, CHAI, etc); I don't count contributions from industry, just the single most relevant academic field; I don't count non-technical academic contributions; and a high discount is applied to academic work. For the sake of argument I've set the discount very high: a unit of adjacent academic work is said to be 80% less effective than a unit of explicit AGI work. The models rely on my priors; customise them before drawing conclusions (see ‘Parameters’ below).

### **A cursory check of the model**

The above implies that there should be a lot of mainstream work with alignment implications - maybe as much as EA produces. A systematic study would be a big undertaking, but can we at least find examples? Yes:  $aix * AIXI$  (2000), a theoretically optimal RL agent.

- Gödel machines (2003), the limit case of verified self-improvement.
- Inverse reinforcement learning (2004 - 2016). A limited but fruitful model for thinking about value learning.
- Various forms of Imitation learning
- Active learning, particularly TAMER (2009) and active reward learning (2014).
- Info-theoretic measures of control like empowerment (2005).
- Adversarial training (2015). As used in AI Safety Debate.
- Wooldridge on the game-theoretic / social choice agent foundations of AI.
- Existence proof for the short/long-term overlap: The Stanford “Center for AI Safety” is a good example. Zero mention of AGI or alignment while working on many of the de facto topics.

By comparison, how much does EA safety produce? In Larks' exhaustive annual round-up of EA safety work in 2019, he identified about 50 paper-sized chunks (not counting MIRI's private efforts). Of them, both CAIS and mesa-optimisers seem more significant than the above. Recent years have seen similarly important EA work (e.g. Debate, quantilizers, or the Armstrong/Shah discussion of value learning).

### **What does this change?**

I argue that AIS is less neglected than it seems, because some academic work is related, and academia is enormous. (My confidence interval for the academic

contribution is vast - but I didn't quite manage to zero out the lower bound even by being conservative.) Does this change the cause's priority?

Probably not. Even if the field is bigger than we thought, it's still extremely small relative to the investment in AI capabilities, and highly neglected relative to its importance. The point of the above is to correct your model, to draw attention to other sources of useful work, and to help sharpen a persistent disagreement within EA safety about the role of mid-term safety and academia.

This might change your view of effective interventions within AIS (for instance, ways to bring AGI alignment further within the Overton window), but my model doesn't get you there on its own. A key quantity I don't really discuss is the ratio of capabilities to alignment work. It seems prohibitively hard to reduce capabilities investment. But a large, credible academic field of alignment is one way to replace some work on capabilities.

## Future safety-related work

A naive extrapolation implies that AIS neglectedness will decrease further: in the last 10 years, Safety has moved from the fringe of the internet into the heart of great universities and NGOs. We have momentum: the programme is supported by some of the most influential AI researchers - e.g. Russell, Bengio, Sutskever, Shanahan, Rossi, Selman, McAllester, Pearl, Schmidhuber, Horvitz. (Often only verbal approval.)

In addition, from personal experience, junior academics are much more favourable towards alignment and want to work on it.

Lastly: Intuitively, the economic incentive to solve AGI-safety-like problems scales as capabilities increase and as mid-term problems draw attention. Ordinary legal liability disincentivises all the sub-existential risks. (The incentive may not scale properly, from a longtermist perspective, but the direction seems good.)

If this continues, then even the EA bet on direct AGI alignment could be totally outstripped by normal academic incentives (prestige, social proof, herding around the agendas of top researchers).

A cool forecasting competition is currently running on a related question.

This argument depends on our luck holding, and moreover, on people (e.g. me) not naively announcing victory and so discouraging investment. But to the extent that you trust the trend, this should affect your prioritisation of AI safety, since its expected neglectedness is a great deal smaller.

## Parameters

- Your probability of prosaic AGI (i.e. where we get there by just scaling up black-box algorithms). Whether it's possible to align prosaic AGI.

Your probability that agent foundations is the only way to promote real alignment.

- The percentage of mainstream work which is relevant to AGI alignment. Subsumes the capabilities/safety overlap and the short/long term safety overlap. The idea of a continuous discount on work adjacent to alignment would be misguided if there were really two classes of safety problem, short- and long-term, and if short-term work had negligible impact on the long-term problems. The relevance would then be near 0.
- The above is extremely sensitive to your forecast for AGI. Given very short timelines, you should focus on other things than climbing up through academia, even if you think it's generally well-suited to this task; conversely, if you think we have 100 years, then you can have pretty strong views on academic inadequacy and still agree that their impact will be substantial.

### Caveats, future work

- To estimate academia fairly, you'd need a more complicated model, involving second-order effects like availability of senior researchers, policy influence, opportunity to spread ideas to students and colleagues, funding. That is, academia has extremely clear paths to global impact. But since academia is stronger on the second order, omitting it doesn't hurt my lower-bound argument.
- If you have an extremely negative view of academia's efficiency, then the above may not move you much. (See for instance, the dramatically diminishing return on inputs in mature fields like physics.)
- A question which deserves a post of its own is: "How often do scientists inadvertently solve a problem?" (The general form - "how often does seemingly unrelated work help? Provide crucial help?" - seems trivial: many solutions are helped by seemingly unrelated prior work.) I'm relying on the overlap parameters to cover the effect of "actually trying to solve the problem", but this might not be apt. Maybe average academia is to research as the average charity is to impact: maybe directly targeting impact is that important.
- I haven't thought much about potential harms from academic alignment work. Short-termists crowding out long-termists and a lack of attention to info hazards might be two.
- Intellectual impact is not linear in people. Also, the above treats all (non-EA) academic institutions as equally conducive to safety work, which is not true.
- Even more caveats.

# Serious science fiction

Gavin

2020-07-17

You live in a hard sci-fi story.

science,

social,

moral / social

software

Dick tried to do social and moral but failed

Le Guin did social without much of the others

Vinge does software and social

Tchaikovsky science and

Egan

Chiang

Banks all about moral and social

# Against the Culture

Gavin

2020-08-17

{% include banks/links.md %}

Liberalism is a technology for preventing civil war. It was forged in the fires of Hell – the horrors of the endless seventeenth century religious wars... from the burning wreckage, we drew forth this amazing piece of alien machinery. A machine that, when tuned just right, let people live together peacefully without doing the “kill people for being Protestant” thing. Popular historical strategies for dealing with differences have included: brutally enforced conformity, brutally efficient genocide, and making sure to keep the alien machine tuned really really carefully.

– Scott Alexander

INTERVIEWER: ... the Ships and Minds of the Culture, its great AIs: their outrageous names, their dangerous senses of humour. Is this what gods would actually be like? BANKS: If we're lucky.

The two worst omissions from sci-fi are social development and software development. In his *Culture* series Banks covers the first so memorably, so thrillingly, that the series is a permanent touchstone for me, even though each individual book is actually not that strong. The Culture is actually different from us - even though underneath their society revs our great alien machine, liberalism unbound.

Ada Palmer calls it “social science fiction”, focussing on soft technology and cultural progress rather than rigorous physics and cool gadgets. A pencil is technology. But so is liberalism, in some sense. Banks was a determinist, and so denied the dichotomy: the technology creates the society. “Space minus scarcity implies anarchism.”

How can anarchism be stable, though? Banks doesn't say it is: instead it's metastable. If your society is a matter of degree, if its only hard tenet is “do what you like if it doesn't hurt anyone”, and if you don't need specialisation of labour, you can get away with decentralisation.

Almost all of the books center on Special Circumstances, the tiny military intelligence branch of the Culture. They are the least typical members of the

Culture, often officially not members. They are central because their lives lend themselves to exciting fiction and because the tensions of the culture are most obvious there (see Critiques below).

As a novelist and a standard Scottish radical, Banks was incapable of writing a pure utopia: no story without problems. Every book has its greys and queasiness: there are three or four critiques of the Culture in the books, sometimes given by Culture citizens. He mostly solves this by having the antagonists be clearly much, much worse than the ultra-democratic luxury altruists. And I shouldn't overegg his pessimism: he is able, after all, to see a world with technological fixes to social organisation and individual suffering.

Banks' world is an achievement: he maintains narrative tension despite having supremely powerful protagonists, post-scarcity bliss, and post-Singularity rationality and benevolence. Culture floats free of physical constraints, and unlike most sci-fi (most fiction) he actually imagines us into that possibility. Where philosophy and art are almost the only big things left to do.

An easy formula is that you wouldn't want to live in *anyone* else's utopia. This is too neat: I would be Culture if it was offered. It just falls short of the real radical optimum.

## What's so good about it?

Briefly, nothing and nobody in the Culture is exploited. It is essentially an automated civilisation in its manufacturing processes, with human labour restricted to something indistinguishable from play, or a hobby. No machine is exploited, either; the idea here being that any job can be automated in such a way as to ensure that it can be done by a machine well below the level of potential consciousness...

- Post-scarcity. No greed.
- Post-disease.
- Post-death.
- Post-gender, post-race.
- No admin.
- Sustainable bliss. Fun recognised as the deep moral value it is.
- Full morphological freedom
- Ability to estimate consciousness and value and so promote it.
- Full positive and negative liberty
- Massive devaluation of ascribed identity in favour of achieved.
- Benevolent decentralised overlords. Unmitigated consent as iron law.
- Freedom of movement and exit. Partial identification (“80% Culture”).
- Almost negligible crime, and so no criminal justice, and so no dedicated police or bureaucracy 1.
- Almost no internal politics.

Banks calls this anarchism, but it is equally a technocracy, or a million little

benevolent dictatorships.

## Critiques of the Culture

### 1. Horza: the Culture as tutelage, just a game

CNN: In the Culture's post-scarcity society, where no one needs for anything, you're removing a lot of the struggle around everyday life. Is that not removing the point of life itself? BANKS: I think a lot of the struggle is kind of pointless and is in itself boring. The struggle for existence for most people most of the time, especially in a post-agricultural, industrial society, is a bit of a grind. People have to work very hard and awfully long hours for not a great deal of money: if you don't, you get virtually nothing. Life's not much fun, frankly, so I'd quite happily trade in that struggle.

while they may be fun, hobbies are also at some level always frivolous. They cannot give meaning to a life, precisely because they are optional. You could just stop doing it, and nothing would change, it would make no difference, which is to say, it wouldn't matter.

– Heath

The humans are not the protagonists. Even when the books seem to have a human protagonist, doing large serious things, they are actually the agent of an AI. (Zakalwe is one of the only exceptions, because he can do immoral things the Minds don't want to.) "The Minds in the Culture don't need the humans, and yet the humans need to be needed." (I think only a small number of humans need to be needed - or, only a small number of them need it enough to forgo the many comforts. Most people do not live on this scale. It's still a fine critique.)

The projects the humans take on risk inauthenticity. Almost anything they do, a machine could do better. What can you do? You can order the Mind to not catch you if you fall from the cliff you're climbing-just-because; you can delete the backups of your mind so that you are actually risking. You can also just leave the Culture and rejoin some old-fashioned, unfree "strongly evaluative" civ. The alternative is to evangelise freedom by joining Contact.

One of Banks' protagonists is anti-Culture. The boring version of his critique is that he dislikes machines ruling humans - their enemies are on the side of life - "boring, old-fashioned, biological life". But the real point is that the Culture's all very well for the actively questing, protagonist Minds, but terrible for its lesser subjects, because nothing in their life is truly serious, counterfactual, functional. Horza thinks you need struggle, ultimate meaning, grand narrative. He sides with the Idirans because at least it's an ethos. (As always, the Culture partially assimilates this critique: one of them names itself after him and so his objection.)

There are objective limits to their egalitarianism (e.g. the artificial-intelligence Ships are straightforwardly superior to their organic charges):

Look at these humans! How could such glacial slowness even be called life? An age could pass, virtual empires rise and fall in the time they took to open their mouths to utter some new inanity!

and even the Ships have a status ladder:

there was a small amount of vicarious glamour associated with it; guarding the weirdo, letting it roam wherever it wanted, but maintaining the fraternal vigilance that such an enormously powerful craft espousing such an eccentric credo patently merited.

### **1b. Scruton: the Culture as idiot meaninglessness**

the fulfilment of wishes is both one of civilisation's most powerful drives and arguably one of its highest functions; we wish to live longer, we wish to live more comfortably, we wish to live with less anxiety and more enjoyment, less ignorance and more knowledge than our ancestors did

– Banks

Roger Scruton can always be counted upon to piss in the beer: he believes that ubiquitous wonder and joy is impossible, or would make us swinish idiots, “a kind of postmodern individual” he doesn’t want to be seated next to at a dinner party:

Everything deep in us depends upon our mortal condition, and while we can solve our problems and live in peace with our neighbours we can do so only through compromise and sacrifice. We are not, and cannot be, the kind of posthuman cyborgs that rejoice in eternal life, if life it is... The soul-less optimism of the transhumanists reminds us that we should be gloomy, since our happiness depends on it.

Banks shares this worry to some extent; see (3) below for how his utopians are not really posthuman. “Luckily”, the Culture citizens are not in fact free of suffering. For instance, Ulver is incredibly annoying, annoyed, and shallow, and is the personification of Scruton’s critique. (Admittedly she is a teenager, but why would we need teenagers?)

Critique (1) is about the sad need for authenticity and agency, not just freedom and fun.

(1a) is the (so-called) paradox of freedom: if you can do anything, if there are no fixed points, then your choice isn’t meaningful.

### **2. Heath: the Culture as replicator**

The only desire the Culture could not satisfy from within itself was one common to both the descendants of its original human stock and the machines they had (at however great a remove) brought into

being: the urge not to feel useless. The Culture's sole justification for the relatively unworried, hedonistic life its population enjoyed was its good works; the secular evangelism of the Contact Section... actually interfering (overtly or covertly) in the historical processes of those other cultures.

The very best essay on the Culture is 'Why the Culture Wins' by Joseph Heath. He notes that if we view the Culture from outside, as a replicator, then of course it needs a moral mission, of course it has to have interventionist compassion as a core value: that's how such a highly moral meme can spread itself. Despite being small and atypical, Contact is the heart of the Culture, its deep justification for itself.

what does it mean to say that Contact arranges things so that the "good guys" win? It means that it interferes on the side that shares the same values as the Culture. There is more at stake here than just individual freedom. For instance, with the development of technology, every society eventually has to decide how to recognize machine intelligence, and to decide whether AIs should be granted full legal and moral personhood. The Culture, naturally, has a view on this question, but that's because the Culture is run by a benevolent technocracy of intelligent machines... This is what gives the Culture its virulence – at a fundamental level, it exists only to reproduce itself. It has no other purpose.

– Heath

The claim is: A society freed from the need to pay attention to reality, to produce, will be given over to intense memetic drift and competition.

From a certain perspective, the Culture is not all that different from Star Trek's Borg. The difference is that Banks tricks the reader into, in effect, sympathizing with the Borg. Indeed, his sly suggestion is that we – those of us living in modern, liberal societies – are a part of the Borg.

– Heath

You can view any successful process as an amoral replicator. The real question is whether its instances have value - more value than the alternatives. Well...

### **3. Culture humans as insufficiently posthuman**

I praised the level of social development in the books. But his humans aren't *radically* different from us. Critique (1) and (2) only hurt because human nature in the Culture is still recognisable as our nature.

Culture citizens tend to not want to live more than 400 years for some reason. (Sheer deepity: "*death is regarded as part of life, and nothing, including the universe, lasts forever. It is seen as bad manners to try and pretend that death*

*is somehow not natural; instead death is seen as giving shape to life.”*) I don’t expect this to be true.

They are not beyond suffering and competitive stress: note Ulver’s whining and tantrums. Grief is common, sometimes lasting a century. They don’t take wild-animal suffering seriously.

Both humans and Minds are still status-conscious. A Ship which has too high a turnover of human population loses face among its peers. There are celebrities, and renowned artists, debutantes and limited capacity events. (“*Not being spoken to, not being invited to parties, finding sarcastic anonymous articles and stories about yourself in the information network; these are the normal forms of manner-enforcement*”)

The humans are clever but not superintelligent. Why, when there is so much profound superintelligent material to understand?

Mostly humans remain with a pretty conservative tetrapod shape, despite their morphological freedom. This implies a lot of conformism and herding (even just our heavy constraints on attractiveness).

Banks has the books’ distinction between biological humans and AIs coming after a period in which there was no distinction, where the humans were more integrated and cyberised. It’s not clear why you’d return.

Some of them still have conservative ideas of meaning. “*The Culture’s sole justification for the relatively unworried, hedonistic life its population enjoyed was its good works.*” As if pleasure and freedom needed further justification! This mania for authenticity is realistic but painful. The desire to experience and create things seems to me to be a complete substitute for the desire for status, for feeling useful, for validation. But to put it mildly this isn’t universal yet.

The tech is mostly stagnant, apparently because of physical limits.

To some extent the above legacies could be Banks leaning on existing human traits in order to write good relatable fiction, rather than his own philosophy. But not wholly or mostly.

#### **4. The Culture as (partial) reverse alignment?**

AI alignment is the process of making sure that your systems act for the benefit of people, even when the systems are much more powerful than humanity. In Banks’ books, there’s some evidence that the reverse has happened, of aligning humans to Minds.

There is a weird absence of resentment and power-seeking among the posthumans. As we know them, humans constantly chafe under government; the lighter the oppression, the more obvious the chafing. Only a small number of humans are driven to lead and orchestrate large moral projects. And we see almost no unilateral human folly: we don’t see any doomed human coups, for instance.

The example I've spotted is the Culture language being engineered to produce certain philosophies in its speakers. (Sapir-Whorf is false for natural language, but who knows what can be done when you have control over both the processor and the instructions?) Maybe by the 8th millennium they've already done all they need to; maybe they are beyond man-machine politics because the humans were subtly shaped until there were no more tensions that needed politics.

Now, humans as we are are sorely in need of shaping, and the Minds are mostly far more moral than us. However, there are marks of subterfuge which you wouldn't want to see in a utopia.

This critique is not particularly biting, since humans remain awkward and recalcitrant, and need to be bribed on the occasions where a Mind wants them to do something. There are still some awful passions: murderous or sex-mad.

## 5. The Culture as imperialism

the Culture doesn't actively encourage immigration; it looks too much like a disguised form of colonialism. Contact's preferred methods are intended to help other civilisations develop their own potential as a whole, and are designed to neither leech away their best and brightest, nor turn such civilisations into miniature versions of the Culture.

For completeness I should mention this, though I think it is misguided at best. The Contact division presume to convert illiberal (e.g. torture porn) civilisations to utilitarianism (mostly via diplomacy and positive incentives rather than through their overwhelming gunboats). They also police large parts of known space, preventing as many conflicts as they can.

There are people who, reacting to the terrible parts of our history of foreign intervention, reject all such intervention. (They sometimes prohibit even nonviolent intervention.)

This is slightly blunted by the above passage: it's only nonsuffering and tolerance that they enforce on others, rather than hedonism, polymorphism, atheism, anarchism. (OK, they also stomp carbon chauvinism.)

The Prime Directive of *Star Trek* is a fictional example of this. They're supposed to ignore non-space-faring civilisations, up to and including letting them die in natural cataclysms. However, the writers and the characters reject it all of the time: it's violated in dozens of episodes, generally in a way that strikes me as blatantly the right thing to do. This is because the principle sounds better than it is.

To be fair I should reconstruct an actual argument:

1. There are no single true values, or anyway we don't know them. (Philosophy is too weak, or we are.)

2. If we don't have the true values, we cannot justify imposing our values on others.
3. So we cannot justify imposing our values on others.
4. So do not intervene when values conflict.

(This doesn't stop us intervening in a society when its own values are violated by external forces, like natural risks or other invaders.)

Premise 2 is the weak one. We know of many things which are universally bad for mammals. What we lack is a precise statement of the good. But that torture or genocide is bad is not very culturally mediated!

There are difficult forms of the concern though:

- What does the Absolute Liberal do with intolerant enemies?
- What can you do with people who don't want freedom, tolerance, management, diversity?

Critique (2) is related to this, but I think that's just the descriptive form.

## Vulnerabilities

The Culture is mostly shown as more powerful than its foes, able to adapt and match whatever threat, in almost all cases without even compromising its own values. Ships produce Ships, so any big Ship could reconstruct the whole civilisation given time. How then could the Culture fail?

### 1. Running out of moral patients

<h3>Morality, big and small</h3>

<div>

```
Ordinary morality holds that saving one life from one dramatic hazard once is <a href="#">
<!-- -->
What larger things could you aim at? It could just be life-saving on a grand scale, or t
<br><br>
Epic morality provides firm and serious meaning to those conducting it. It is a fine sub
</div>
```

the Culture accepts, generally, that questions such as 'What is the meaning of life?' are themselves meaningless... we make our own meanings, whether we like it or not.

Banks gives the Culture an ultimate meaning (roughly, reducing suffering and promoting freedom), but it's contingent: it needs to keep finding people to help. Assuming that no faction goes totalitarian and starts engineering new terrible societies, a crisis of legitimacy should eventually come. (Though since they don't even fully cover one galaxy by the end of the timeline, it'll be a while.)

Even then, there's no real prospect of a successful human revolt. So nothing left to do except Sublime, chase other realities.

Alan Jacobs' standard sniffy zero-sum critique ("a society without internal struggles will need always to generate external ones") is unfair and anyway unnecessary for this to be a vulnerability.

## 2. Meeting a stronger replicator.

Banks' world contains Hegemonising Swarms: collections of self-replicating matter, not sentient, not creative, just very good at destruction, reproduction, and travel. Swarms are the logical extreme of an illiberal foe: one with no values, only reproduction. Watts' *Blindsight* contains a formidable sort. (All of the very powerful agents in Banks' books are sentient: he tacitly assumes, against Watts, that consciousness is adaptive.)

The Culture spends most of its resources on recreation. (The Minds are not vigilant, spending large amounts of time in an intellectual opium haze.) Even if we grant Heath's cultural evolution point, Contact is a vastly expensive and slow method of reproduction compared to a Swarm. Then there's the unseriousness of everyone (Human spy: "He'd thought about saying, Well, actually I was in [the secret service], kind of a spy, really, and I know lots of secret codes and stuff...")

Most of the potential sentient threats to the Culture have "sublimed" (dematerialised); the Culture is an aberration, kept in reality by their civilising mission. (There are maybe a dozen "Involved" civilisations on their level.) But there's nothing to stop another civ with a conflicting moral and more focus on fast spread also refusing to go buddha.

## 3. Space Balkanisation

the forces of repression need to win every time, the progressive elements need only triumph once.

– Banks

Each Ship is a nation-state. The anarchic collection of mobile states works because there is a strong vetting process for new minds, which prunes away the psychos and megalomaniacs, and provides a bedrock of strong Millian consequentialism in nearly all Minds. (One of the few rogue elements in the series, the *Attitude Adjuster*, is still a good utilitarian with a horror of killing, and is utterly overcome by guilt at the deaths it causes while trying to end a systematic torture culture.)

Even so, the Culture has no mechanism for preventing schisms, besides the meta one of 1) basic shared consequentialism, 2) not limiting its members enough to make it necessary to schism. The path to failure is ideological drift -> civil war or recursive schisms -> lack of coordination -> military loss.

The ship training process is imperfect, and there are still schisms and hot conflicts in the Culture among the aligned Minds. In his early theoretical notes, he talks

about the difficult process of becoming the Culture: overcoming many intolerant local minima, and phrases the Culture as what happens when your hegemony is so total that you don't need to enforce it anymore. But conflict still lurks out there, even when you're beyond economic and strategic concerns.

They seem to have no central authentication or strategists, only temporary committees. The military, and even the secret service, are fully decentralised, subvertible by any single high-status rogue.

Almost none of the humans in *Excession* are actually wholeheartedly 'in' the Culture; instead there are only factions: the Elench, the AhForgetIt, and an allophilic Culture diplomat who ends up defecting entirely. Maybe the Culture is constituted of people who don't feel totally in it, but who recognise that everyone else is worse.

Big reasons for hope:

- The Culture's ship fertility rate could easily be high enough to indefinitely replace these defections.
- The ex-Culture factions depicted still co-operate a lot. Their philosophical differences do not extend to deep casus-belli questions. (Exception: the Elench's pathological curiosity and touchy-feeliness.)
- Placing no barriers on separatism, while retaining ideological agreement on harm, lets the Culture seem much smaller than it is while maintaining an otherwise-threatening extent.
- Status markets. There is still positional scarcity, and reputational risks, which prevents most bad behaviour and wireheading. "one of the many tiny but significant and painful ways a Ship could lose face amongst its peers was through a higher than average crew turn-over rate"

## Misc

- Subliming is a really, really bad plot device. To stop recursive self-improvement and first-mover advantage from making his galaxy boring, Banks has all of the really powerful civs voluntarily dematerialise for mysterious spiritual reasons. Even in a soft world (with e.g. basically no energy constraints), this breaks fictional belief. Maybe he had plans to make it less bad which he didn't get around to; maybe it would have tied it to the extra-dimensional beings of *Excession*.
- The Minds are not improving much; ancient ones orchestrate many of the grand successful space operas. This is odd.
- The Minds are funny. They are addicted to super videogames. They gossip, and they plot, and they can dislike each other. They do all this a billion times faster than us, in amusing cryptographic ways, but they remain comprehensible and likeable superintelligences. We should expect even aligned superintelligences to be much stranger than this: mind design space is too large, and our concepts too small, for it to be otherwise.

- The Culture are against terraforming - an odd apparent bit of bioconservative ideology. But this seems to be mostly a matter of efficiency: artificial habitats are much more efficient.
- If there is anything to the neocolonialism / ‘liberal hegemony’ suspicion, the sad fact remains that it’s a less bad hegemony than the others.
- There are no religions in the Culture, not even relatively rational ones like simulationism (shown in a different Banks book) or panpsychism or deism. With so much free time, alongside sports, art, and philosophy, I expect humans to get into unprecedently odd metaphysics.
- Minds have strong emotions (e.g. the ROU Killing Time’s kamikaze fury).
- They have brain uploads, but they’re mostly just in storage and are greatly outnumbered by embodied people.
- The Minds run incredibly detailed simulations of terrible situations; there’s no attention to whether this is morally risky.

I know it’s all nonsense, but you’ve got to admit it’s impressive nonsense.

– Banks

## See also

- Banks, ‘A Few Notes on the Culture’
- Heath, ‘Why the Culture Wins’
- Yudkowsky, ‘The amputation of destiny’
- Sandifer, ‘Cultural Marxism 1: *Consider Phlebas*’
- Jacobs, ‘The ambiguous utopia of Iain M Banks’
- *The Age of Em* is the hardest social science fiction I know, albeit written as nonfiction. What do our best nonphysical theories imply?

{% include banks/foots.html %}

# Blogging is dead, long live sites

Gavin

2020-04-27

{% include sites/links\_and\_style.html %}

Blogging peaked in 2009; I was there, just.

{% include sites/trends.md %}

Writing a ‘web log’ was mostly social: like a public diary; as if everyone was a hyper-local newspaper columnist with a letters page. “Here’s what I’ve been up to”; “here’s my reaction to what Dubya just said”; “I just remembered a thing”; “here’s why (a)theists are dumb”. More about process than product.

Half a dozen people used to comment on everything. There were various Spheres, where amateurs and professionals of various sorts thought out loud and gossiped and bitched. Famous economics professors followed my obscure fumbling posts, inexplicably.

It had an economy, thousands of people making a living off it (or anyway someone making money off it). That racket is still there, but the clever or young people long ago moved to Youtube, Insta, podcasts, and newsletters.

Blogs were supplanted by centralised social media. This was maybe because they’re more effective for broadcasting and harvesting status, and because no-one there is aiming for more than ephemerality. (When was the last time you looked at what anyone said on Facebook last year?) In the transition period, your Twitter or Tumblr page got called a microblog.<sup>2</sup> But people moved on, making the original unit superfluous.

My point: ‘Blogging’ has been used for both short-term indie punditry/self-expression, and long-term indie creative/intellectual work. The first is now on social media. The second lives on: I learned more from these personal sites than I did in three stints at university. Many of these sites are called blogs, but I say leave the word to the first thing.

An overlooked fact: the internet dies off at an astounding rate. The average link stops working after about two and a half years. Not only was blogging reactive and local; it was also mortal.

{% include sites/2x2.html %}

Why prefer the bottom-right? Why not write ephemera, or for oneself only?

No binding reason: just if you want to do something big; if your ego or your morals demand it; if you want to seed more than a one-time flurry of agreement, disagreement, indifference, impressions. The rest of this piece is about the second column.

## Essays vs blogposts

blogging is not a form of writing... Blogging is an activity that is so distinct from the experience of writing that it should be called something else altogether. One does not write a blog post except in the sense that one "writes" a shopping list or a business plan...

Blogging, in my experience, reduces writing to the short-term effects you have on your readers and they have on you. You try to have an immediate, essentially real-time impact on the discourse, which makes it much more like speech than writing. . . .

– Thomas Basbøll

So, when I'm in a poncey mood, I say I don't write blogposts - they're essays. This ain't no blog, it's a site! Basbøll uses "Writing" for the real deal, to be backwards compatible with the likes of Roland Barthes. But this is the most confusing possible name for it.

In a way it's funny to set up essays as a superior substitute for blogging / social media musing, since "essay" (*attempt*) was itself self-consciously inferior to big tedious monographs from the start. ("The essay - or microtreatise.")

But never mind terminology. The imitable Gwern aims for "long content", work updated continuously for decades, living, growing piece by piece into magnificence:

how do you write a personal site with the long-term in mind? We live most of our lives in the future, and the actuarial tables give me until the 2070–2080s... What sort of writing could you create if you worked on it (be it ever so rarely) for the next 60 years? What could you do if you started now?...

what would constitute Long Content as opposed to the existing culture of Short Content?...

the pages will persist through time, and they will gradually improve over time. But a truly Long Now approach would be to make them be improved by time—make them more valuable the more time passes.

It's not about being pompous or pretending to have timeless wisdom; it's the attempt to do things that become more and more amazing, which are worth keeping updated, worth living up to.

e.g. Depending on the field, a PhD might involve reading 100 - 400 papers, doing a thousand hours of asynchronous, unpredictable Innovative Work, then writing about 5 papers. Minus the admin, the teaching load, the mandatory courses, etc: call it 6000 hours. If you did this for 5 hours a week, say on a Sunday afternoon after waking late and having brunch and ambling about, you could do something of comparable scale over about 20 years: with no financial implications, no sweat, no mental breakdown. *While working full-time on other things*, and with my life 30% gone already, I could do 5-10 things this hard, just with suitable long-sighted use of weekends. 3

(As it happens even serious academic work is surprisingly volatile; around half the links in the average academic paper die within a decade. There are often alternative ways to recover the target document, but not always.)

(As it happens I think most PhDs don't have much impact on the world: they are read by say 4 other people, and maybe *should* not be read by many more than that. But that's good: instead I can do a thousand bits, each their own contribution to the future of all things.)

If you're reading this, you probably have a lot of energy, up to 10 big tickets. What do you want to spend them on?

```
{% include sites/examples.html %}
```

## Independent and academic scholarship

I haven't said anything about where to do this work. (In what institution.)

A lot of the sites I list in the accordion above are by part-time autodidacts, or retired scholars. I suppose this is because the incentives in academia are so often towards either small publons or giant monographs, each of which are set in stone once done. (Unless they are grossly flawed enough to trigger academia's slow, dumb immune system.) (Or my search process is biased towards lone wolves.)

But the average academic work is more lasting than the average internet work. But it isn't only durability we're after. But it's also more rigorous than the average internet work.

Robin Hanson has spotted a trend among independent scholars, a systematic bias against rigour, and so against durability.

over time amateurs blow their lead by focusing less and relying on easier, more direct methods. They rely more on informal conversation as analysis method, they prefer personal connections over open competitions in choosing people, and they rely more on a perceived consensus among a smaller group of fellow enthusiasts. As a result, their contributions just don't appeal as widely or as long.

Take Hanson himself: he has about 100 academic publications, two big books, and something like 3000 blog posts. Which will be his biggest contribution in

the end?

Maybe tenured academics are the people best placed to do long content: lots of time, lots of connections, some pressure towards rigour and communicability. But you should be able to do it outside uni, if you're wary.

Think tanks are the usual way to be a full-time intellectual outside academia. But there are innovations that could enable group reinforcement and dialectic on a wider scale, for the many great part-timers: Researchers.one is the fullest version so far. Also The Winnower; The II. Preprint servers and post-publication review.

## See also

- Long content is really uncommon. Even great internet writing with a view to the long term (e.g. Eliezer Yudkowsky's sustained braindump of 2007 to 2009) is never updated, when its problems are found at all. This should worry us, since it implies it's hard to do. Maybe few people have stable enough goals and interests to do this, or just enough time.
- The situation may be even worse in open source software, with projects overwhelmingly dead by 6 months old.
- Basbøll, 'What is blogging?'
- Gwern, 'Long Content'
- Stock and flow (2010)
- This has something to do with Digital gardens, but those are just an intermediate public phase between raw notes and final essays. Alive though.
- Link rot

{% include sites/foots.md %}

# Marvellous measures

Gavin

2020-06-29

{% include measure/links.md %}

[In 1690] there were no standard “fixed points”, namely phenomena that could be used as thermometric benchmarks because they were known to take place always at the same temperature. Without credible fixed points it was impossible to create any meaningful temperature scale, and without shared fixed points used by all makers of thermometers, there was little hope of making a standardized scale.

— Hasok Chang

Anything can be measured. If a thing can be observed in any way at all, it lends itself to some type of measurement method. No matter how “fuzzy” the measurement is, it’s still a measurement if it tells you more than you knew before. And those very things most likely to be seen as immeasurable are, virtually always, solved by relatively simple measurement methods.

— Douglas Hubbard

For a very long time, we could only look, only feel. You checked claims against your own forearm, your own foot, a scrap of crop, or you paid a rare savant to count for you.

How can you do more than this? You need a way to compare two things in terms of something other than a person. A fixed point, a reference object is handy, so you can resolve disputes and check your instrument. Eventually, if you’re smart, you get a unit; if you’re extremely smart you get a whole system of convertible units, tying together almost the whole world.

These moments, these initial quantifications, are an incredible thing: the initiation of a new domain into science; the ability to use the world around us to track and change what’s inside us. Here are my favourite occasions on which a subjective, ineffable thing was suddenly partially outside of us, less contestable.

We leap from the angle of a stick in the sun to the size of the whole planet; from some crappy artefacts in your data to the very composition of the sun. (I guess

Eratosthenes' doesn't count, since length and geometry were already admitted to be objective, just unknown.)

When can you use relations between numbers to talk about relations between objects?

Quantification is ordinal: you go from the subjective and qualitative, to the objective but ordinal, to the objective and numeric, to the objective and precise. I'm mostly talking about the first shift: from essentially subjective to suddenly not. I also ignore whether something is objective or merely universally intersubjective.

```
{% include measure/accord.md %}
```

## Small Physics

- Joule
- [https://en.wikipedia.org/wiki/Mole\\_\(unit\)](https://en.wikipedia.org/wiki/Mole_(unit))
- rads

## Big physics

- Erastosthenes and the circumference.
- [https://en.wikipedia.org/wiki/Anthropic\\_units](https://en.wikipedia.org/wiki/Anthropic_units)
- Temperature. Fahrenheit [https://en.wikipedia.org/wiki/Timeline\\_of\\_temperature\\_and\\_pressure\\_measures](https://en.wikipedia.org/wiki/Timeline_of_temperature_and_pressure_measures)
- Beaufort wind force scale
- Richter. log
- decibel. log
- megaton
- Mohs. Ratio, not interval.

## Really big physics

Length was one of our first conquests: the Akkadians had a reference length. AU, light-year, parsec are basically just logical consequences of that.

[https://en.wikipedia.org/wiki/Spacetime#Spacetime\\_interval](https://en.wikipedia.org/wiki/Spacetime#Spacetime_interval) [https://en.wikipedia.org/wiki/Comoving\\_and\\_p](https://en.wikipedia.org/wiki/Comoving_and_p)

## Time

- Sidereal time. Beyond solar position. <https://en.wikipedia.org/wiki/Hour#History>  
[https://en.wikipedia.org/wiki/History\\_of\\_calendars](https://en.wikipedia.org/wiki/History_of_calendars)
- Geological time
- Tree ring width

## Information

- The bit. Maybe the most profound of all, since it not only transformed our society but also turned out to be a fundamental concept in the best physics

we have. (NB: To an usual degree, the technical meaning of information and the ordinary meaning are very divergent. Maybe we'll reunite them one day.)

## Uncertainty

We can guess the distribution of almost anything. (The ideas involved are so simple it makes you wonder why we waited til the Manhattan Project to have them, but many ideas seem that way after a Martian has had them for you.)

- Resampling
- Monte Carlo (MC) simulation

## Life

- Urea. Stoichiometry
- [https://en.wikipedia.org/wiki/Mark\\_and\\_recapture](https://en.wikipedia.org/wiki/Mark_and_recapture)
- [https://en.wikipedia.org/wiki/Category:Ecological\\_techniques](https://en.wikipedia.org/wiki/Category:Ecological_techniques)

## Idea quality

- factor analysis, path analysis, multidimensional scaling
- Inter-rater reliability
- Reliability. (The famous one, Cronbach's alpha, is notably rigid and biased downward.)
- Loss function. A model (a hypothesis) is better if it has a lower loss on new data.

## Psychophysics

<https://en.wikipedia.org/wiki/Category:Psychophysics> [https://en.wikipedia.org/wiki/Just-noticeable\\_difference](https://en.wikipedia.org/wiki/Just-noticeable_difference) . e.g. the Scoville scale

In the method of just-noticeable differences, the experimenter made small variations in the physical intensity of a stimulus, and determined when the subject detected the stimulus, or noticed that it had changed; In the method of constant stimuli, the experimenter presented two stimuli that differed slightly in magnitude, and determined when the subject detected a difference between them; In the method of adjustment, the subject – not the experimenter – varied the magnitude of the stimulus, and indicated when it (or a change) became detectable.

gLMS <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3501107/>

Wilhelm Wundt on feeling <https://www.ocf.berkeley.edu/~jfkahlstrom/ConsciousnessWeb/Introspection/Psychph>

## Health

- Number needed to treat.
- QALY.
- Micromort
- [https://en.wikipedia.org/wiki/Life\\_table](https://en.wikipedia.org/wiki/Life_table)

## Psychometrics

[https://en.wikipedia.org/wiki/Epworth\\_Sleepiness\\_Scale](https://en.wikipedia.org/wiki/Epworth_Sleepiness_Scale) Beck depression inventory  
Personality Moral psychology <https://blogs.lse.ac.uk/politicsandpolicy/happiness-tests-wellbeing/>

## Sexuality

[https://en.wikipedia.org/wiki/Klein\\_Sexual\\_Orientation\\_Grid](https://en.wikipedia.org/wiki/Klein_Sexual_Orientation_Grid)

## Value

Economists are often accused of believing that everything - health, happiness, life itself - can be measured in money. What we actually believe is even odder. We believe that everything can be measured in anything.

— David Friedman

- (Ordinal) utility
- QALY
- The dollar PPP [https://en.wikipedia.org/wiki/Big\\_Mac\\_Index](https://en.wikipedia.org/wiki/Big_Mac_Index)  
[https://en.wikipedia.org/wiki/Holmes\\_and\\_Rahe\\_stress\\_scale](https://en.wikipedia.org/wiki/Holmes_and_Rahe_stress_scale)  
<http://www.overcomingbias.com/2020/06/the-value-of-life.html>

## Anything

Dollar Subjective probability

```
<h3>Bad measures</h3>
<div>
    That's the happy bit.<br><br>
    Many, maybe most measures fail to approximate their target.<br><br>
    All measures become bad measures if someone has an incentive to fool them.
    Anyone who works in an organisation will have encountered some bad measures, often ones
    * SLOC
    * Subjective wine quality. Sulphites are disparaged.
</div>
<h3>Future measures</h3>
<div>
    * <a href="{{pain}}">Real pain scales</a><br>
```

```
* <a href="{{quale}}>All qualia</a>.<br>
  * The (cardinal) util.
</div>
<h3>Against measurement</h3>
<div>
  * <a href="{{rain}}>Some romantics</a> view quantification (and explanation more genera
  <!-- -->
  * "Lies, damned lies, and statistics": it is easy to make up numbers. But it is even eas
</div>
```

## See also

- Metrology: the science of measure
- Measurement theory: the mathematics of when number relations can express object relations.

# Things I Use

Gavin

2020-04-11

No affiliate links, because I am lazy. Prices given are what I paid, usually during a sale.

{% include stuff/links.md %}

## Health

- Bowflex SelectTech 552 dumbbell (£180). By my desk; much harder to avoid than the gym. 15 dumbbells in one. I love how little space they take, and the design is extremely satisfying. You can get by with one. *Expected lifespan:* 15 years? *Cost per year:* £12.
- Power rack and barbells (£1000 for own gear, or £30 / month gym membership). Barbells give me big structural and mental changes. Commercial gyms are fine but you can do far better in environment and cost if you have room for your own gear, or know somebody who does. I was lucky enough to have a mate with a free garage. They also keep their value really well, so resale should be roughly the same as initial cost, which might mean that you actually save relative to gym. *Expected lifespan:* Bar should be 10+ years. Rack maybe 20 years. Plates N/A. *Cost per year:* £50 if you have space / £250 if renting a space with two other people.

```
<h3>Advantages of garage gym over commercial gym:</h3>
<div>
  <ul>
    <li>Privacy</li>
    <li>No time wasted commuting</li>
    <li>No time wasted queueing</li>
    <li>Can workout in anything, save on gym gear</li>
    <li>Music of choice</li>
    <li>No idiots dropping 150kg from head height at 110dB.</li>
    <li>No disease</li>
    <li>Open 24h </li>
    <li>Plausibly cheaper than public gym (< £250/year if you can split with friends).</li>
  </ul>
</div>
```

```

<h3>Buyer's guide</h3>
<div>
  I got <a href="{{rack}}>a set</a> second-hand. For iron plates, £2 / kg is a good price
  <!-- -->
  Garages are pretty costly in cities. Outside of London you can find them for <£80 a month
</div>
<h3>Other gear</h3>
<div>
  There's no limit to the amount of gimmicks you could buy, from 30x markup powders to be
  <!-- -->
  I'm trying out <a href="{{shoe}}>weightlifting shoes</a> for barbells (£70). Feel nice
  <i>Expected lifespan</i>: 3 years?
</div>



- Dahon Espresso D24 folding bike (£320 on the Cycle To Work scheme, down from £530). Full size wheels and front suspension: fine for offroad. Folding means you can take it on trains, subways and car boots despite the size. The wheels it comes with are ridiculously thick, but you can get road ones. Probably saves me about 100 hours a year walking, and also gives me joy. I loved the old version, which had a bike pump hidden in the seat column. Expected lifespan: 10 years light use with £100 annual service. Cost per year: £120 per year, amortised.
- Powdered greens (£25 for 100 days). I like leafy veg, but the cost and the low shelf-life makes me eat it less than I want. I mix 10g of this into my morning shakes and feel virtuous at least. Wait for it to be 40% off. Cost per year: £80
- Zinc acetate lozenges (£10 for 30). There's some evidence that keeping particular ligands of zinc in your mouth when you have a cold speeds up your recovery, maybe by a full day. Many other viruses also perish under zincky attention, so they're out of stock as of Spring 2020. Cost per year: £5
- Sleep mask (£8). I slept poorly after I moved to the city, mostly because of ambient light. Now I don't! You want one which curves away from your eyelashes and isn't too hot. This is excellent. Expected lifespan: 2 years? Cost per year: £4
- Oxymetazoline (£8 for months). Never mind the Vicks or the steam bath. This clears your nose in seconds; makes sleeping when ill much easier. Cost per year: £8
- Anyone can book a winter flu jab for about £10. IIRC, in expectation this should save you about 12h of misery / lost work. Expected lifespan: 1 year Cost per year: £10
- Either stannous fluoride (£3 for 2 months) or hydroxyapatite toothpaste (£3 for 2 months). (The normal kind of toothpaste has the less effective

```

sodium fluoride.) I haven't seen one with both, so might be good to alternate. *Expected lifespan:* 2 months. *Cost per year:* £15 over normal paste.

- Blood tests are cheap now! (£30 or so.) Even if you don't feel bad, there's very little reason not to check, say every two years. Vitamin D and iron are a good start; "subclinical" deficiencies of them are common. I found I had slightly low ferritin, and this was such a cheap thing to fix. This service signs you up to a biannual subscription, but you can just cancel after the first one. *Expected lifespan:* 2 years *Cost per year:* £15

## Food

- Queal shakes (£1 per meal). I was skeptical at first: "complete meals" (artificial food) seem procrustean (assuming dietary science is finished) and joyless. But it's based on oats and soy protein. This dissolves much better than Huel and has loads of flavours. I pad it out with rice protein powder and so usually get about 5 shakes out of a bag. *Cost per year:* Same as a solid breakfast.
- MSG powder (£4 for 200 meals). Average vegan food is good but just needs a lot more fat and umami. I get the latter from a sprinkling of magic dusts (MSG and nutritional yeast). There is no good evidence that it has any negative health effects, and in fact it's a little healthier than table salt (less sodium).
- You want one good chef's knife. (You hurt yourself less with a really sharp knife since its motion is more predictable.) I just bought a random £25 one in Tesco and it is excellent. *Expected lifespan:* 3 years if you sharpen it *Cost per year:* £10
- Fastbake breadmaker (£50). British supermarket bread is crap. This makes 900g of warm, chewy, custom bread for about 55p. Chuck in poppy seeds or linseed or nuts for 10p, done. Wholemeal is trickier, needs a little bit of vitamin C powder. Power is maybe 5p. *Expected lifespan:* 4 years? *Cost per year:* roughly the same as shop bread.

## Productivity

- Filco Majestouch mechanical keyboard (£99). Feels amazing, built like a tank. The noise takes some getting used to. I went for Red switches (quieter but also less tactile). I don't need a numpad but *maybe* you do. (PS: you have to love Japanese web design, as long as you don't have to rely on it.) *Expected lifespan:* 10 years? Or never if I get into soldering. *Cost per year:* £10
- Sublime Text (£35). I do basically all of my writing and coding in this editor. Wonderful search, regex, markdown, and build options. Packages

for everything. There are languages that really need IDEs (Java comes to mind), but I don't write in em. You can get it free, but I wanted to support it. I hear VS Code is even better but I am content. *Expected lifespan:* Forever.

- This blog is hosted on Netlify, they are amazing and free for small fry.
- Wire up your laptop for calls (£5). Who knows how much of social difficulties are due to people disliking latency on your calls? Replace the cable every couple years just in case.
- Bose QuietComfort 25 (£150). Being able to turn down noise is a superpower: actual focus. They even made working in an open-plan office intermittently tolerable. They don't work *that* well on conversations, but are excellent for humming appliances, engine roar, wind. Lasted 3.5 years so far. These are the wired ones because I resented paying an extra £100 for a transponder; I've had to replace the cable twice, 2 x £7, and probably about £20 of batteries. *Expected lifespan:* I guess 5 years? *Cost per year:* £30
- Nicotine lozenges (8p a day). Better than caffeine. Vaping is cheaper but riskier and restricted in many locations. Takes a little while to get used to the mild burning. This is the only thing on this list with real risks. *DO NOT EAT A WHOLE ONE WHEN YOU START; start with 0.5mg or less and don't swallow it. Cost per year:* £30
- Amazing Marvin is the nicest to-do list ever. It's programmable and supports dozens of different productivity systems, recurring reminders, timers, whatever. Lifetime subscription is pretty cheap during Xmas sales too (£150). *Expected lifespan:* Forever. (10 years)
- 45W Corn Light (£20). I don't have SAD, but during winter I noticed a little bump in mood and energy from hanging up 3 of these very bright blue LED clusters. Each is about 400W equivalent in terms of halogen bulbs. *Expected lifespan:* Probably 10 years.
- A big plastic timer (£15). Useful for cooking and for remembering that the pomodoro work technique exists. The original brand is ridiculously expensive. *Expected lifespan:* 10 years. *Cost per year:* £1.50
- ThinkPad Carbon X1 laptop with customs maxed out (£1600). Light, fast, beautiful, runs Linux without a peep. m.2 drive is worth every penny. 1 *Expected lifespan:* 6 years. *Cost per year:* £250
- I don't know if it counts as productivity, since I spend about an hour a day playing with it, but Roam is the best personal knowledge base software I've seen. Text, maths, code, images, bidirectional links, single-copy imports... It promises to unify me across decades. (My blogs also do this, but only for the top 1% of thoughts.) Workflowy and Notion are a tree: Roam is

the awesome power of a graph, which is what thoughts are like. Currently free, soon to be pricey.

## Travel

- Berghaus Freeflow 35+8 backpack (£80). This has a clever mechanism at the back to shrink and grow the volume by 25%, and also a harness to leave a gap between your back and the bag, preventing deathly hike sweats. On extra small mode it fits even stingy Ryanair airline cabin requirements (there's some optical illusion about you wearing your cabin bag on your back, I've never been bothered about it in 43L mode. (If you wear two jumpers and a jacket just for passing through the gate, 35L is two weeks' basics, no cabin bag.) I've had this for 8 years, maybe 100 difficult trips including long haul airports and 1km mountains, and it's fine. *Expected lifespan:* 'Lifetime guarantee'. (20 years?) *Cost per year:* £4
- Moto G7 Power (£160). I resisted getting a smartphone for 8 years. I still think it's a huge threat to productivity, and a privacy disaster. But for travelling it is a massive help: boarding passes, Maps, taxis, translation, mobile data. Also allows me to replace my ereader, my GPS, my trips to ATMs, my camera, my printer. Group chats have been relatively useful already. The new UK Railcard is app-only too. This has the largest battery life on the market (26h of low-res video playback), and is cheap and good. *Expected lifespan:* 4 years. *Cost per year:* £40

## Services, Security

- KeePassX password manager (£0). Works on every platform: Linux, Win, Mac, Android. Probably saves a few minutes a week and a lot of mental overhead. See here for why you want this.
- Protonmail is free and actually secure.
- Private Internet Access VPN (£50 per year). VPNs are imperfect, but they help mitigate a few different problems (IP tracking, unencrypted traffic, ISP logs, public wifi spoofing, geo-locking, app requests). PIA got a subpoena for their logs and they came up clean. Again, see here.
- Pi-hole ad blocker (£25). Stops ads at the source, for every device in your house at once. *Expected lifespan:* 5 years.
- Vanguard ETFs. One of the most surprising facts is that automatic index funds outperform "actively managed" (paying a finance person) ones, after you subtract their fees. Vanguard are the original and are among the lowest fees, about 0.15% of your return. I use a variant of the Simplicity Portfolio and rebalance every 6 months. You may be amused to hear that they are "communist". Above, I said that only nicotine has any real risk -

but these are a layer of abstraction over the stock market, so obviously be careful.

- Focusmate (£3 a month). I work from home a lot, and this lets me force myself to have arse in chair by 8:30am. That's worth it alone. One friend thinks it makes him 20% more productive on top of that; I'd say 5%. (Come join me in the EA room!) *Expected lifespan*: 1 month.
- Emergency backup bank account (Free). A couple times in the last decade I've been locked out of my account due to a false-positive for fraud, or lost my bank card. To be able to get to the bank / to work during the day (or four) this takes to resolve, I have a backup bank card with about £100 on it. You can also just stash a little cash in your house, but this is more general than that. Free, takes maybe one hour including the appointment some UK places make you do. *Expected lifespan*: Forever.

## Fun

- KS Miami bluetooth speaker (£20). Surprisingly good bass; makes watching things on a laptop much less dreadful. Good battery life too. *Expected lifespan*: 4 years. *Cost per year*: £5
  - Tailored socks. One of my favourite possessions. I have socks which *actually* fit for the first time. (The creator took 6 measurements!) It was a gift, but I would probably pay £30 if I was rich. *Expected lifespan*: 3 years.
  - Two actually nice shirts (£40). There are a lot of weddings in my life at the moment. And besides that it's nice to surprise people once in a while. *Expected lifespan*: 3 years. *Cost per year*: £15
  - Fairy lights make all rooms nicer, any time of year (£20 for loads).
- 

## Cost-effectiveness

Rather than just telling you their cost, I should say how much good they do per pound. Ignoring the free ones, which you should just go and get now, I think the best are:

1. Vanguard ETFs. Negative cost, and they're hard to beat on returns/fee unless you're full-time Finance. NaN:1
2. Sleep mask. Massively improved sleep quality, without having to alter the room, close the windows, whatever. 100:1.
3. Dumbbells. A cheap gym membership is £150 a year; using these a couple times a week for 2 years means I've saved hundreds of pounds and dozens of hours commuting. They should last 15 years, so maybe total 30:1. (During

the present lockdown, with gyms closed, the dumbbells get a temporary massive boost too.)

4. Meal shakes once a day. Saves money (if a lunch would otherwise be £4) and time. Also a handy automatic prepper store. 10:1.
  5. Mechanical keyboard. Assuming this decreases my RSI risk by 1%, it will have paid off 10 times over. But also in comfort and fun alone. 10:1
- 

Why write this? One of the big bottlenecks to improving your life is just knowing that it's possible to improve a given part. For some reason people don't share their data on this, probably a reaction against vulgar consumerism.

## See also

- 1000 nerds
- Scott Alexander
- Rob Wiblin
- Sam Bowman
- Mark Xu
- Alexey Guzey
- Jose Ricon
- Peter McCluskey
- Philip Storry
- Various (2012)
- Various (2020)
- Rosie Campbell
- Michelle Hutchinson
- Arden Koehler
- Louis being mean
- Spencer's joy
- 100 bookish types

# Stimulant tolerance, or, the tears of things

Gavin

2020-10-07

```
{% include phone_img.html %} {% assign ath = "https://jissn.biomedcentral.com/articles/10.1186/1550-2783-7-5" %} {% assign sig = "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2738587/#R20" %} {% assign brain = "https://brainlabs.me/" %} {% assign rog = "https://www.nature.com/articles/npp201071" %} {% assign kar = "https://www.ncbi.nlm.nih.gov/books/NBK430790/" %} {% assign gwern = "https://www.gwern.net/Nootropics#caffeine" %} {% assign down = "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3437321/" %} {% assign rog2 = "https://link.springer.com/article/10.1007%2Fs00213-012-2889-4" %} {% assign fluke = "https://slatestarcodex.com/2019/08/19/maybe-your-zoloft-stopped-working-because-a-liver-fluke-tried-to-turn-your-nth-great-grandmother-into-a-zombie/" %} {% assign redd = "https://www.reddit.com/r/Supplements/comments/ls7caq/t" %}
```

Withdrawal from caffeine causes... headache, fatigue, decreased energy/activeness, decreased alertness, drowsiness, decreased contentedness, depressed mood, difficulty concentrating, irritability, and feeling foggy/not clearheaded... abstinence from low doses, such as about one small cup of coffee per day, also produced symptoms of withdrawal.

— Karima et al (2009)

Despite 100 years of psychopharmacological research, the extent to which caffeine consumption benefits human functioning remains unclear.

— Rogers et al (2013)

You wake at 7am and, as usual, immediately put the coffee on.

Consider two (extreme) scenarios for a quickly-eliminated, chronic-use substance like coffee:

4

In one case, the caffeine lifts you above your previous performance and then resets over the day. In the other, the morning caffeine undoes a withdrawal,

returning you to your previous performance, but then makes you *worse* than before by the end of the day.

The worry is that chronic use of caffeine creates a problem and then masks it by associating itself with the relief of those symptoms. (This might further be disguised if acute caffeine use is like the left hand graph.)

(Again, that right-hand graph is not realistic. In reality the graphs will cross; the rest of this is just wrangling over where the crossing is. Does caffeine actually raise average performance?)

The theoretical reason to expect that chronic use looks like the right-hand graph is “downregulation”: the body foils most attempts to permanently increase almost anything to do with the brain. Then there’s a bunch of empirical studies that attempt to measure whether chronic effects are net positive.

- *strong withdrawal reversal hypothesis*: the cognitive effects of chronic caffeine use are not net positive.
- *beneficial naivety hypothesis*: the cognitive effects of caffeine use are net positive in naive users.

I take this seriously enough that I quit caffeine. But if withdrawal reversal *and* beneficial naivety are both true, it implies that we should instead *cycle* caffeine (use on one day, then take a break). But then what’s the cycle length that avoids harm?

---

Could billions of people really do something every day and not notice it has no net effect? Could science fail to discover or communicate this for decades, despite the experiments being cheap and safe? (...)

---

### Causal graphs

The sceptical hypothesis is that  $|b+c| > |a|$ .

### What could caffeine help with?

To see the overall effect of caffeine, we need to distinguish closely related effects:

- Mental stamina (how long you go)
- Subjective energy (how hyped you feel)
- Motivation (how able to start you are)
- Vigilance (how much you focus)
- Working memory (how well you keep current tasks in mind)
- Recall speed (how quickly you remember past things)
- Recall accuracy (how well you remember) 1

We could bundle all of these up into “productivity”, but that’s extremely difficult to measure.

Gwern, who has looked into this more than me, continues to take it, since he assumes the motivational effects are net positive:

For me, my problems tend to be more about akrasia and energy and not getting things done, so even if a stimulant comes with a little cost to long-term memory, it’s still useful for me.

This just raises the question of whether the motivation effect is real or a reversal though. (It might be that not all of caffeine’s effects get blocked over time; for instance the athletic gains.) I suppose motivation is easier to subjectively check than cognition.

### **As always, cost-benefit**

This post suggests that there might be no cognitive benefit. But that’s only one part of the coffee phenomenon:

- Money cost
- Anxiety
- Maybe some harm to long-term memory
- But if you like the taste, then it’s probably worth it (you might be confusing withdrawal relief with flavour though)

### **How fast does tolerance build?**

Karima:

caffeine withdrawal occurred after as little as three days of caffeine exposure

This suggests that if you’re cycling it, you could take it two days on, n days off. (With n somewhere between 2 and 30.)

(In practice, people only seem to use it daily or never, but I don’t know if this bimodal thing is mostly biological.)

### **Literature**

This question - whether 80% of all adults are fooling themselves - is surprisingly little studied, and the ones that have been done include lots of useless n=20 studies.

There was a cluster of work in the early 00s, mostly confirming or consistent with withdrawal reversal, but very little since.

Keywords: “withdrawal reversal”, “net [beneficial] effects of chronic administration”.

## The real deal

I found one good recent study with strong pre-hoc controls for confounding: Rogers (2013), n=369, blinded:

The terribly named “Non-low” mean “non-consumer or low-consumer of caffeine”.

Overall, the high caffeine users are worse than the non-lows when each are given placebo, and are not notably better than the non-lows when both are given caffeine. This is consistent with withdrawal reversal.

## Mere physiology

Sigmon et al (2009) :

There was almost no evidence for net effects of chronic caffeine administration on these measures [cerebral blood flow velocity, EEG, and subjective effects].

## Alertness

Rogers et al (2010):

Caffeine did not increase alertness in [low-intake] participants. With frequent consumption, substantial tolerance develops to the anxiogenic effect of caffeine, even in genetically susceptible individuals, but no net benefit for alertness is gained, as caffeine abstinence reduces alertness and consumption merely returns it to baseline.

James (2014) is methodology showing how difficult we find it to get rid of the withdrawal confounder (for just one: caffeine crosses the placenta, so there may be very few humans who are *truly* caffeine naive!).

## Design for a self-experiment

Caffeine metabolism is mediated by several genes we know about 3. In effect this means that we need to produce a few different estimates, one for each relevant genotype; it could well be that you’re one of the lucky ones which the above genetically naive studies gloss over.

If you don’t have your DNA sequence, or if you distrust the maturity of caffeine genetics, as you should, then you want to run an experiment:

## Outcomes

1. Cognition
2. Motivation
3. Subjective wellbeing

Leave productivity out of it for now.

## Aims

1. Work out if the chronic gain is larger than the withdrawal harm *for you*.
2. If it is, then work out the optimal cycle period *for you*.

## Protocol

1. Quit caffeine. Zero intake for 3 weeks
2. Use Cambridge Brain Sciences to get a caffeine-naive baseline. Say at least a month of that.
3. Use 100mg powder at the same time every day (more ergogenic, precise dose).
4. Add 100mg theanine (since we're interested in the best case rather than the isolated effect of caff).
5. Daily tests for say another month.

(For later blinded experiments, get a friend to produce a coded pillbox of alternating 100mg caffeine powder and 100mg cornflour. Add these to something strong like 100ml orange juice to mask the extreme bitterness. Pulpy juice should help cover textures too. Quinine is similar )

## Other stimulants

The above reasoning about downregulation applies to most other stimulants: nicotine, modafinil, etc. Nicotine apparently interacts with caffeine, so you'd want to do a clean univariate experiment for each.

## See also

- Gwern, who gave me the idea
- Lovely collection of links and questions.

{% include stims/foots.html %}

# Interactive Artificial Intelligence CDT

Gavin

2020-08-24

# Are we moral?

Gavin

2020-09-25

{% include nihil/links.md %}

Some ways we might fail to be moral

- viciousness: intentionally failing by some moral standard
- akrasia: wanting to be moral but failing because of willpower
- apathy: Not even trying
- moral error: wanting to be moral, and trying, but failing because of lack of knowledge (empirical or normative)
- philosophical moral nihilism: “there are no moral properties to make actions succeed”
- psychological moral nihilism: “people don’t have any moral intentions”. Trying, but seemingly moral actions are actually disguised egoism.

The last two are the grandest problems and discussion of them isn’t centralised. Fault in ontology, or a fault in fundamental psychology.

## Metaethical nihilism

It might be that there are no moral properties: “we all always fail to be moral, because there is no way to succeed”. Ontology

<https://medium.com/@tommycrow/what-is-your-meta-ethical-position-c27939810985>

## Descriptive egoism

And/or it might be that people don’t have moral intentions, despite appearances.

(called “psychological egoism” in the literature, but this is a bad name: it connotes a subjective absence of altruism intentions, which is straightforwardly false.)

<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0378.2009.00379.x>

**Soares' Stamp Collector thought experiment (which is psychologically very convincing)**

**Egoism makes good predictions**

sure, but you have unsatisfactory ontology: we want altruism/egoism dichotomy to distinguish certain classes of behavior which are useful; we don't really care about "true motives".

e.g. microeconomics

**Evolutionary argument**

My confidence is almost atheoretical: there are just (a small number of) people whose actions are very difficult to reconcile with stamp-collector-register logic, e.g. Irena Sendler or many people who resisted their local social incentives and were killed for it. (Limited to atheists for the purposes of a clean counterargument.)

Ethical egoism is different again: the idea that it is moral to not be altruistic.

<h3>Confusing real examples</h3>

<div>

Ludwig Wittgenstein. Do harm for loved ones. We are appalled, but know we would do the s

Jeff Bezos. Vast power to help, omit to help.

<blockquote>We had to think about it, because one could build hospitals with that money

<http://www.eonline.com/news/39358/abba-s-billion-dollar-rejection>

Agnetha . Vast power to help, omit to help.

</div>

# Favourite social theorists

Gavin

2020-09-19

{% include nonstem/links.md %}

[Say] I have... a young white male student, politically-correct, who says: "I am only a bourgeois white male, I can't speak." ... I say to them: "Why not develop a certain degree of rage against the history that has written such an abject script for you that you are silenced?" ... investigate what it is that silences you, rather than take this very deterministic position – since my skin colour is this, since my sex is this, I cannot speak... if you make it your task not only to learn what is going on through language, but also at the same time through a historical critique of your position, then you will have earned the right to criticize, to be heard. When you take the position of not doing your homework - "I will not criticize because of my accident of birth, the historical accident" - that is the much more pernicious position.

— Gayatri Spivak

Ninety percent of everything is crud.

&#8213; Theodore Sturgeon

I was a romantic teen, so I did a philosophy degree. (Philosophy, it turns out, includes a lot of the most relentlessly unromantic, counterintuitive, and indeed anti-human writers in the world.)

But I didn't focus. Instead I went to lectures in sociology, politics, gender studies, history, anthropology, and "Area Studies". When I read economics I only got on with the self-styled "heterodoxy". I fell in with the kinds of people who think they have done intellectual work by calling something neoliberalism, who mock attempts at objective analysis of human behaviour merely because it is really hard and often done poorly.

Later, when it finally came time to go and Do Something, I got into analysis, evaluation, and decision theory. I even learned some maths. I realised that my misgivings about e.g. economic analysis were about a strawman; that e.g. genetics was at that very moment transforming our understanding of human prehistory and human nature.

Somewhere along the way of becoming a scientist, without realising it, I had fallen into the culture war, the terrible idiot descendent of the 90s science wars. This usually involves writing off many entire fields as intellectually bankrupt. I should really have avoided this, since I had deeply valued these things just a few years before. But this memplex is very good at making smart people do stupid things.

Queer theory is not all of LGBT studies. Critical race theory is not all of Black studies. Nor are all of queer theory or critical race theory discreditable. (One moment which jolted me out of my silly lazy see-sawing was the above Spivak quote, calling people in to think for themselves regardless of identity. For those who don't follow academic gossip, Spivak is a descendent of the arch-villain Derrida.) The degree of stereotyping and weak-manning on each side of the culture war is really, really disheartening.

You have to ignore some things; human thought is too large not to. But you don't have to sneer, and you certainly don't have to generalise like a fool. All kinds of thinkers know what you can lose by thresholding a continuous variable; the others have a more verbal (and less accretive) way of putting it.

So right now I'm enjoying going back to the start and mining derided thinkers for insights. With charity and a spot of strong misreading, this is generally great fun.

This is not to say everything's equally fruitful, rigorous, making progress at an equal rate or value. I fully expect to hit another level of the dialectic someday, when methods for detecting degenerative research programmes improve and I can objectively ignore. But in the meantime it looks like few fields have only one programme, and few have only nonsense ones.

---

## Sociology

- Pierre Bourdieu. People think he's one of the pseuds just because he writes really badly. But he was never part of the edgelord cafe set, he reinvented PCA, and was the best empiricist in the field for a long time.
- Kieran Healy. His twitter is one of my top 10.
- Collins

## Cultural anthropology

- Ernest Gellner.
- David Graeber. An exception in this list, because he says false things fairly often and more generally filters evidence poorly.
- Latour

I'm not going to cheat and include any of the many awesome physical anthropologists like John Hawks.

## Law

- Tom Bingham

## LGBT studies

- John d'Emilio.

Sadly, Hite and Kinsey are not worth reading except for historical interest.

## Gender studies

- Serano?
- Rich?
- Eve Sedgwick?

## Critical theory / philosophy

- Rorty
- Nancy Fraser
- Ian Hacking's The Social Construction of What?

## See also

- Ben Hoffman on extracting models of epistemic decay from Baudrillard
- My discussion of the anti-science turn in American cultural anthropology, and my resolve to read it anyway.

# Using courts for algorithmic fairness

Gavin

2021-03-16

{% assign lgfo = “<https://arxiv.org/abs/2009.11677>” %}

I got some fan mail on a paper I helped with last year!

Dear Mr. Leech I am a current senior at [High School], and I am currently researching predicting United States district court case verdicts. After reading your article, I had several questions I was wondering if you could find the time to answer: 1) Is it possible to apply the LGFO algorithm to determining verdicts in court cases? How would you go about doing that? 2) How were you able to bypass potential bias when creating this algorithm? 3) While this algorithm can be applied to a binary classification, how could you potentially expand it to help in the training of the model? I would greatly appreciate any answers you could provide.

---

I don't think any algorithm exists that can make verdicts on its own. And LGFO isn't intended to decide court cases. Instead it uses data from courts as a way of working out how to balance the many kinds of fairness, for any classifier which is making predictions about social input.

You might have heard that there are lots of ways of putting fairness into mathematical form, and that many of them contradict each other. You literally cannot satisfy them all. How then do we decide how unfair something is? How do we decide how much each type of fairness counts?

Our system solves this as follows:

- a. A human picks a set of fairness definitions
- b. A human gives the algorithm a set of past cases, along with the damages awarded in each case.
- c. LGFO works out how much weight to give each kind of fairness, and so produces a classifier which is as fair as possible, if we trust the legal system to know this relatively well.

It does this by assuming that the amount of money awarded in a case scales closely with the unfairness of that case.

Now, your questions:

- 1) Is it possible to apply the LGFO algorithm to determining verdicts in court cases?** It gives you a general classifier, so nothing technically stops you applying it to verdicts, or to recidivism predictions. But it wasn't developed for this and would only accept simple numerical inputs (like the defendant's age). I wouldn't use it in courts in its current form: it's like a prototype which would need a lot of work to customise for justice applications, because the stakes are so high and a bad system could really harm people.
- 2) How were you able to bypass potential bias when creating this algorithm?** All systems are biased, in the weak sense that you can't satisfy all fairness measures at the same time. The advantage of LGFO is that it limits the bias to be only as severe as the legal system it uses for input, in particular the civil tort system of your country. The bias of most other systems is not so limited: the user makes all kinds of decisions (e.g. the definition of fairness, the weight to give each definition, the thresholds at which the classifier flips) which could be much more biased. This isn't perfect, but at least the law is a partially democratic process. It's hard to see how to do better than this.
- 3) How could you potentially expand it to help in the training of the model?** There are lots of ways to extend it. One really easy way to turn a binary classifier into a multi-class classifier (e.g. from one which says "Hot / Cold" to "Hot / Warm / Lukewarm / Cold / Freezing") is to use "one-vs-rest": basically you train 5 binary classifiers and take the one with strongest confidence as the answer.

I hope your project goes well!

# Why worry about future AI?

Gavin

2021-03-21

```
{% assign lp = "https://online-optimizer.appspot.com/?model=builtin:default.mod"
%} {% assign eh = "https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg"
%} {% assign gcoin = "https://www.google.com/search?q=0.5%5E5" %} {% assign ggun = "https://www.google.com/search?q=1+%2F+%286*6%29" %} {% assign arm = "https://www.youtube.com/watch?v=WLXuZtWoRcE" %} {% assign ny = "https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai" %} {% assign zi = "https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/" %} {% assign robo = "https://arxiv.org/ftp/arxiv/papers/1507/1507.03518.pdf" %} {% assign gpt = "https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html" %} {% assign ipcc = "https://www.givewell.org/shallow/climate-change/extreme-risks#The_problem_Risk_of_worse-than-expected_impacts" %} {% assign agi = "http://gcrinstitute.org/2020-survey-of-artificial-general-intelligence-projects-for-ethics-risk-and-policy/" %} {% assign oai = "https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secrective-reality/" %} {% assign wino = "https://en.wikipedia.org/wiki/Winograd_Schema_Challenge" %} {% assign glue = "https://gluebenchmark.com/" %} {% assign banana = "https://en.wikipedia.org/wiki/Banana_Massacre" %} {% assign dhp = "https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/" %} {% assign giz = "https://gizmodo.com/for-20-years-the-nuclear-launch-code-at-us-minuteman-si-1473483587" %} {% assign gp = "https://web.eecs.umich.edu/~weimerw/p/weimer-ssbse2013.pdf" %} {% assign lye = "https://www.scientificamerican.com/article/how-hackers-tried-to-add-dangerous-lye-into-a-citys-water-supply/" %} {% assign uk = "https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/" %} {% assign chain = "https://www.guinnessworldrecords.com/news/2016/10/video-watch-australian-daredevils-terrifying-attempt-at-blindfold-chainsaw-worl-449269" %} {% assign russ = "https://www.edge.org/response-detail/26157" %} {% assign omo = "https://selfwaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf" %} {% assign treac = "http://lukemuehlhauser.com/treacherous-turns-in-the-wild/" %} {% assign danaher = "https://philosophicaldisquisitions.blogspot.com/2014/07/bostrom-on-superintelligence-3-doom-and.html" %}
```

## Harm through stupidity

Could AI be a risk to humans? Well it already is:

- Elaine Herzberg was killed by an Uber self-driving car, while walking her bike across a pedestrian crossing. The system couldn't decide if she was a bike or a person, and the switching between these two possibilities confused it. Uber had disabled the Volvo automatic braking system. (It was slowing them down.)
- About one in 100 robot surgeries involve accidents; about 20% of these were what we'd call AI failures (things turning on at the wrong moment, or off, or misinterpreting what it sees). (This seems to be lower than the human rate.)
- Consider also things like the Ziyuan Blowfish, an autonomous Chinese military drone currently under export to the Middle East.

## Harm through intelligence

These systems did harm because they were too stupid to do what we ask (or because the humans deploying it are).

What about a system harming us because it is too smart? Is there any real chance that advanced AI could ruin human potential on a grand scale?

## Argument from caution

We don't know. They don't exist, so we can't study them and work it out. Here's an argument for worrying, even so:

1. It's likely we will make a general AI (AGI) eventually.
2. We don't know when.
3. We don't know if it will be dangerous.
4. We don't know how hard it is to make safe.
5. Not many people are working on this. (<500)
6. So it's probably worth working on.

In particular, your starting guess for  $P(\text{soon \& dangerous \& difficult})$  should be at least 3%.

I just put a number on the risk of this unknown thing. How?

Well, we surveyed 350 mainstream AI researchers in 2017.

- Median P of AGI within a century: 75%
- Median P of "extremely bad" outcome (human extinction, loss of governance, or worse): 5%
- Median P of safety being as hard or harder than capabilities: 75%

If we illicitly multiply these, we get a prior of a 3% chance of catastrophic AGI this century.

This is weak evidence! AI researchers are notoriously bad at predicting AI; they're probably biased in lots of ways (e.g. biased against the idea that what they're working on could be morally wrong; e.g. biased in favour of AGI being soon).

But you should go with 3% until you think about it more than them.

```
<h3>3% is small!</h3>
<div>
    Not really. It's the probability of <a href="{{gcoin}}>5 coin flips</a> all coming up heads.
    Or more pertinently, the p of dying when playing Russian roulette with <a href="{{ggun}}>one bullet</a> in the chamber.
    It's also <a href="{{ipcc}}>roughly the same</a> as the probability of extreme climate change.
    <!-- -->
    Probabilities don't lead to decisions on their own; you need to look at the payoff, which is
</div>
<!-- -->
<h3>High uncertainty is not low probability</h3>
<div>
    The weakness of the evidence means we remain very uncertain - it could be 0.1% to 90%.
    If you are genuinely uncertain about whether there's a landmine in front of you, you don't
</div>
<!-- -->
<h3>Against the null prior</h3>
<div>
    People often act like "things should be treated as 0 probability until we see hard evidence".
    The last year of government failure on COVID should make you think this isn't the right
    It is not possible to have direct evidence yet, so it doesn't make sense to demand it.
</div>
```

---

### Reasons to worry more

#### People are trying hard to build it.

There are 72 public projects with the stated goal of making AGI. Most of them have no chance. But billions of dollars and hundreds of the smartest people in the world are pushing it.

In the study of viruses and bacteria, there's a thing called "Gain of function" research, when you intentionally modify a pathogen to be more lethal or more transmissible. Most AI research is gain of function research.

#### We're getting there.

GPT-3 displays quite a bit of common-sense, an extremely hard open problem. We will probably pass the Turing test within 5 years.

We've already passed a number of other classic benchmarks, including the fiendish Winograd schemas.

OpenAI, the people who made GPT-3, were polled. Their median guess for when AGI was 15 years.

### Indirect evidence of danger

#### The human precedent

There is evidence for intelligence enabling world domination: we did it. (Also through vastly superior co-ordination power.) Chimps are maybe the second-most intelligent species, and they are powerless before us. They exist because we let them.

Another worry from the human case is that we seem to have broken our original “goal”. Evolution optimised us for genetic fitness, but produced a system optimising for fun (including directly anti-fitness fun like birth control and disabling depressants).

Lastly, we are a terrible case study in doing harm without hatred, just incentives. No malevolence needed: chimps are just made of / living among stuff we can use.

The thought is that humans are to chimps as AGI is to humans.

---

#### Intelligence is not wisdom

People sometimes say that it's a nonissue, since any system that is truly intelligent would also be wise, or would know what we meant, or care.

Two counterexamples:

- Human sociopaths: sometimes highly intelligent while lacking any moral sense
- Reinforcement learning algorithms. Their goals (reward function) are completely separate from their intelligence (optimiser / planner).

RL is the most likely current technology to eventually become an AGI. It has a few worrying features: autonomous (no human input as standard), maximising, and with hand-written goals, with <100 variables. i.e. they are told to value only a tiny fraction of the environment.

---

#### Current stupid systems still cheat ingeniously

They come up with ingenious ways to subvert their goals, if that is easier than actually doing the task.

- Coastrunners. An RL bot was given the goal of winning the race as fast as possible. It worked out that actually it could get infinite points if it never finished the race, but just collected these powerups forever.
- A robot was trained to grasp a ball in a virtual environment. This is hard, so instead it learned to pretend to grasp it, by moving its hand in between the ball and the camera. Trying to deceive us.
- GenProg:

A genetic debugging algorithm, evaluated by comparing the program's output to target output stored in text files, learns to delete the target output files and get the program to output nothing. Evaluation metric: "compare youroutput.txt to trustedoutput.txt" Solution: "delete trusted-output.txt, output nothing"

The point of these examples are: We cannot write down exactly what we want. The history of philosophy is the history of failing to perfectly formalise human values. Every moral theory has appalling edge cases, where the neat summary fails.

If we don't write down exactly what we want, then the system will find edge cases. They already do.

The worst kind of cheating is treachery: initially pretending to be aligned, then switching to dangerous behaviour when you can get away with it (for instance, after you've completely entrenched yourself). This seems less likely, since it requires more machinery (two goals, and hiding behaviour, and a second-order policy to decide between them), and requires us to not be able to fully inspect the system we "designed". But we can't fully inspect our current best systems, and it too has already been observed in a system not designed for deceit.

---

### **We can't even make groups of humans (e.g. corporations) do the right thing.**

No one at an oil company loves pollution, or hates nature. They just strong incentives to pollute. Also have strong incentives to stop any process which stops them ("regulatory capture").

We've maybe gotten a bit better at aligning them: corporations mostly don't murder thousands of strikers anymore.

We should expect AI to be worse. The parts of a corporation, humans, all have human values. Almost of them have hard limits on how much harm they will do. Corporations have whistleblowers and internal dissent (e.g. Google employees got them to pull out of military AI contracts).

(Governments are much the same; it wasn't the United Fruit Company that fired the rifles.)

---

### **Most goals are not helpful.**

Look around your room. Imagine a random thing being changed. Your chair becomes 3 inches shorter or taller; your fridge turns upside down; your windows turn green, whatever.

Humans want some crazy things (e.g. to cut fruit out of their own mouths with a chainsaw).

But for most possible goals, no one has ever wanted them

(“Replace the air in this room with xenon gas” “Replace the air in this room with freon gas” “Replace the air in this room with radon gas...”)

i.e. Human-friendly goals are a small fraction of possible goals. So without strong targeting, a given goal will not be good for us.

We currently do not have the ability to specify our goals very well, and the systems aren't very good at working them out from observing us.

Argument:

1. Hand-written goal specifications usually omit important variables
2. Omitted variables are often set to extreme values.
3. So hand-written specs will often set important things to (undesirably) extreme states.

(To convince yourself of (2), have a go at this linear programming app, looking at the “model overview” tab.)

---

### **Society is insecure**

```
<h3>When will the first anonymous internet billionaire be?</h3>
<div>
```

```
    This has already happened. The anonymous creator of bitcoin holds 1 million BTC, and the
    So we see that immense value can be created - just using programming + internet + writing
```

```
<!-- -->
```

```
    Once you have a billion dollars and no morals, there's not a lot you can't do.
```

```
</div>
```

Our societies are increasingly vulnerable to hacking. Last month someone tried to remotely poison a Florida city's water supply. A few years ago, large parts of Ukraine's power grid were shut down, just as a civil war erupted.

The American nuclear launch code was, for 20 years, “0000000”. What else is currently wide open?

## Maximisers are risky

1. Intelligence and benevolence are distinct. So an AGI with unfriendly goals is possible.
2. A maximiser will probably have dangerous intermediate goals: resource acquisition, self-defence, resistance to goal changes.
3. So a maximising AGI will default to dangerous behaviour. And it might be that you only get one chance to load your values into it.

A corporation is a profit maximiser, and this is probably part of why they do bad stuff.

Again, all of the best current systems are maximisers.

---

## The mess of society

A.I. hasn't yet had its Hiroshima moment; it's also unclear how such a decentralized & multipurpose field would or could respond to one. It may be impossible to align the behavior of tens of thousands of researchers with diverse motives, backgrounds, funders, & contexts, in a quickly evolving area.

– Matthew Hutson

All of the above is how hard it is to solve a *subproblem* of AI safety: 1 AI with 1 human. Other problems we need to at least partly solve:

- Deep mathematical confusion
- Philosophical baggage (can't teach values if you can't agree on them)
- Political economy (arms races to deploy shoddy systems)
- Ordinary software hell (no one writes safe code)
- Massive capabilities : safety funding ratio. 20,000 : 1?
- Treacherous turn
- AI is maybe worse than nukes, climate change, engineered pandemic. Those don't follow you, don't react to your countermeasures.

And huge questions I didn't even mention:

- “Intelligence explosion”
  - Do future people matter?
  - Will AGI be conscious?
  - What is the right decision theory?
  - How much worse is extinction over 99% death?
  - Current leading ideas for solutions (x11)
-

Overall, my guess of this turning out terrible is 15%. One round of Russian roulette.

---

## Sources

Most of the above are other people's ideas.

- Richard Ngo
- Stuart Russell
- Nick Bostrom
- Eliezer Yudkowsky
- Viktoria Krakovna
- Andrew Critch
- Nate Soares
- David Krueger
- Steve Omohundro

## Other links

- DeepMind on real reward hacking
- Long list of real-world ML cheats
- Long list of resources at all levels
- AI Safety Support: Safety coaching charity
- 80,000 Hours prioritise aspiring x-risk people
- My model of the size of AI safety
- Jacob Steinhardt on engineering and safety

# ‘Trompettes de la Mort’ (2005)

Gavin

2021-04-24

```
{% assign yt = "https://www.youtube.com/playlist?list=PLhdvJ4ld4_Ec9wawrK20wue-5Y92aDwqL" %} {% assign lap = "https://en.wikipedia.org/wiki/Rule_of_succession" %} {% assign my = "https://mashable.com/article/myspace-data-loss/?europe=true" %} {% assign print = "https://www.youtube.com/watch?v=eu4M1sIIiVA&list=PLhdvJ4ld4_Ec9wawrK20wue-5Y92aDwqL&index=12" %}
```

Listen

Autumn comes, the trees are naked, I am shaking like a leaf  
Praying for a sudden house-move; a traffic accident; just some relief...  
I only laugh now with permission, do his bidding, got no choice;  
When I speak I say what I'm told to,  
when I speak I don't hear my voice... Even if I could scale the fences, that'd no longer get me out.

– Joe Tucker

My favourite album from 2005 is almost completely ungoogleable. As far as I can tell it's the sole work of Joe Tucker, very slightly better known for National School and better known for his comedy scripts. (His brother plays the cornet in a skit and there's a little bit of extra backing vocals on Track 1, 3, 16.)

It's incredibly *generous*. 20 top-grade melodies - more than one per track, one per minute! - with moving lyrics and bizarrely good production. I'm probably grading on a curve because it's a shoestring labour of love, but I really do love it. (*The Copyright to Life* has played at random in my head about once a month for 10 years.)

Most people's lyrics are very general, but I love the hyperspecific - particular objects, particular idioms, particular moments, particular people. And almost no-one covers the deadening language of the bureaucracies that almost all of us live inside - “transferable skills”, networks, forms, house prices, liabilities. Tucker does, and so brings in the dread and glory of actual life.

Look in the houses, see them glowing tungsten red: Designers, teachers, labourers, brutalist architects. I can hear the breath on your lips, the blood in my head. Someone speaks a house price that dare not be said.

It's normal to not exactly fit your life, and so normal things can express that.

The arrangement is twee: the two most prominent voices are xylophone and Hammond organ. If you're not paying attention you might think the lyrics are twee ("I want to eat vegetables, save the meat for the heavy-hearted"). It also borrows melodies from church, Pink Floyd, Alice Cooper.

---

It's not a concept album but it is a suite, which people often mistake for a concept album. If there is a concept, it's refusing to let life deflate you - even if you find yourself in a little life, alone and ill-conditioned. Or the need to see your life as a story, for all that school and work and ill health demean us, stopping us see ourselves as protagonists. (If you can't manage to live grandly, then fantasise.)

My trade was a whaler - but the whales are no longer biting. I  
retrained in data management and I came back fighting

(At first I thought this was mocking the character, but now I think we should admire his strength, his ability to imbue his life with meaning.)

Trompettes de la mort (trumpets of the dead) are mushrooms. I haven't quite cracked the connection between vegetables and existential resilience, but mushrooms thrive in all sorts of bad conditions. Then there's several songs about the sea which I also haven't related to the social trauma theme or the vegetable theme.

I hear Tucker sing 3 characters: - Track 1, 4, 14 are the narrator (not as shrill) - Track 8 has the only sinister presence, a retired policeman with questionable pastimes. He too is coping. ("high on tea and thoughts of liberty; I'd die for the monarchy"). Authority over others and perversion as two other solutions to the problem of life, besides creativity. - The other tracks seem to be one character.

---

### The protagonist

He tries to be upbeat. He fears the sea, or "the sea". He feels obscurely that school and work have deformed him ("Even if I could scale the fences, that'd no longer get me out.") He gets a job in IT, or maybe insurance. He identifies with Orwell of *Down and Out*: a tourist of poverty. He fantasises to cope.

I'm an in-filling unit of a high performance team I breathe transferable skills, I dream the collective dream I fight from a workstation with a wall partition for a shield Videoconferencing to no man's land, the office is a battlefield

There's a lovely upbeat number about the inventor of CBT, which tells us that he needs that. (CBT is about being able to steer away from false destructive thoughts.) Though *Man Overboard* is pretty clear too:

Looking out to sea - hard to believe that it's going to swallow me.  
Clouds build, seagulls cry. I feel scared, but I will not wipe my eyes.  
Now fetch me those binoculars, there's an island I want to see  
This could be spectacular - feel the silt washing over me. Hold me down:  
I'm not where I'm supposed to be. When I left this town, the girls  
would not speak to me. Hold on tight, grab the sides - careful, we  
might collide Hold on now, mind that wave - my life will not be lived.  
Brave men find my boat, bottom up but still afloat... The luxury  
cruise liner, the cargo ferry: Both look the same at the bottom of  
the sea

It's wise. There's a lot of bad defaults in life, a lot of things to see through: -  
Most writing is hollow and arbitrary and ill-justified. Tucker just points this out  
for tabloids, but the problem is there in almost all papers, academic journals,  
books. - School is very painful for a minority of people, and many of those it  
imprisons don't benefit very much. - People differ, and only some kinds of people  
are served by default kinds of work, socialising, housing, whatever.

He tells me he runs faster, he hits harder, he climbs higher I'm not  
one to argue (and anyway these aren't things that I admire).

The rejection of bourgeois life is one thing. But it's shallow - **as if** just inverting  
something freed you from it. (Similar to the indie snob who dislikes things  
merely because they are popular.) But it's a start.

The amateur astronaut - someone out of their depth, chasing a dream too large  
for them - is a natural image. I've seen half a dozen things use it to great effect.

Her indoors brings me a cup of tea whilst I light the fires The engine  
screams fizzing cacophony as I climb inside Leaving the earth far  
behind of me, danced them into dots - Jupiter, Saturn, and Mercury!  
Wish you were here, love you lots.

---

I think it's my album of 2005 1, and until I dug around in my ancient wma rips  
and stuck it on Youtube, none of you could know it. Which implies that there  
is another album of the year out there, undiscoverable except to the 10 people  
who knew the artist personally. (The album of 2017 is definitely on Bandcamp  
or Youtube, but you'll never find it anyway.)

What else?

The Sunset Tree

Antony & the Johnsons 2

Fiona Apple

Illinois

Kate Bush

Potemkin City Limits

Out-of-State Plates

Black One

Black Sheep Boy

The Campfire Headphase

Sunlandic Twins

LCD Soundsystem 1

Z

Kanye 2

Twin Cinema

Run the Road

Sunset Tree is better, and similarly rich in melody and lyrical detail, but I'd rather listen to Trompettes than any of those.

The death of Myspace left a huge gulf in underground music, up to 50 million tracks lost. Mostly terrible bedroom bands, but surely also works of genius. It won't happen again; Youtube's use of semi-legal fan labour means more or less everything since 2009 is preserved, no matter how obscure.

Listen!

# Do masks work against COVID, at scale?

Gavin

2021-06-19

```
{% assign mxv = "https://www.medrxiv.org/content/10.1101/2021.06.16.21258817v1"
%} {% assign tw = "https://twitter.com/g_leech_/status/1406287131534893059"
%}
```

We have a new preprint!. Here's a full explainer thread.

We seem to be first to use the incredible UMD / Facebook survey of COVID behaviour to look at masks.

Short answer: Yup! 24.6% [6%, 43%] reduction in R the reproduction number, or cases / case.

We also have interesting secondary results

- Voluntary mask wearing started earlier and to a larger extent than previously realised. (64% of the world reported wearing masks by *May 2020*.)
  - We have exactly two examples of noncompliance with mandates.
  - Mask wearing has started falling (about 5% over May 2021) *in countries without fast vaccination campaigns*.
  - Past work used the timing of government mask mandates. You really can't do this, because of the huge voluntary uptake prior to them.
- 

## What's the catch?

- We only use data from last summer, our wearing data is still a proxy (self-reported wearing), and our analysis is observational. See Discussion for lots more.
- Our analysis goes further in the quality of wearing data - 100 times the sample size, with random sampling and post-stratification - geographical scope, the sophistication of our infection model, the incorporation of the uncertainty in epidemiological parameters, and the robustness of our results (123 sensitivity experiments).
- Our analysis begins in May 2020, after some of the earliest mandates, as that's when data first became available.

- Summer 2020 has distinctive features: many regions began with NPIs already active; public behaviour had already changed following the (in)formal instructions of the first wave; and summer months are thought to have lower transmission
- We don't break the effect down by the venue of wearing. We don't look at cultural factors or serious differences in effectiveness of different types of masks. Our analysis is at the national (or US state) level, so we could miss subtler policy effects.

Our definition of 'mask-wearing' isn't stringent: it'd apply to a person who wears a cloth mask, only on public transport, 51% of the time; and to a person who always wears an N95 respirator outside home. So there's scope for more & better wearing, even in regions reporting high levels in our data.

---

Here's a full explainer thread.

Here's the code (end-to-end instructions).

# Highlights from my Gists

Gavin

2021-06-04

```
{% assign dnd = "https://gist.github.com/g-leech/05a106d425fe8477a86acfd0dd1c82d6"
%} {% assign spock = "https://gist.github.com/g-leech/3e3fd37feb23251cf8c245203dfe1f6"
%} {% assign lw = "https://www.lesswrong.com/posts/S3LKfRtYxhjXyWHgN/d-and-d-sci-april-2021-voyages-of-the-gray-swan" %} {% assign pow =
"https://gist.github.com/g-leech/4c4335c665f47ef83a2913cf9a9eb378" %} {% assign alc =
"https://gist.github.com/g-leech/075f47dcf6a66508dbd380b56961b8c8"
%} {% assign tax = "https://gist.github.com/g-leech/4f4b7492bf0c1f34a3a0871c6877b386"
%} {% assign ev = "https://gist.github.com/g-leech/05a106d425fe8477a86acfd0dd1c82d6#file-dnd4-py-L39" %} {% assign covid =
"https://gist.github.com/g-leech/66c76b0a7c623dca0a46a5ef1d1eee2c"
%} {% assign kelly =
"https://gist.github.com/g-leech/01fe74896b12e6d180a1dbc3b77c2fd0"
%} {% assign psy =
"https://gist.github.com/g-leech/80a8b5917ae1fb8baf57c8805c72eee9"
%} {% assign chp2 =
"https://gist.github.com/g-leech/21dbbdebd68d65fb7cfa61b9fbcd508"
%} {% assign chp3 =
"https://gist.github.com/g-leech/1a63f22805053186642a4b93f7dd0f77"
%} {% assign itila =
"http://www.inference.org.uk/mackay/itila/book.html" %} {% assign mask =
"https://gist.github.com/g-leech/5ddbe5bd06ddd3ee2152cd7beb4762b3"
%} {% assign goo =
"https://gist.github.com/g-leech/38a9e40c3cc83ea1f96fcbc0c0fb6657"
%} {% assign q =
"https://gist.github.com/g-leech/8cd4466e4cfa84a8950ef39c5bd813fc"
%} {% assign bda =
"http://www.stat.columbia.edu/~gelman/book/" %}
```

Github Gists are a readable way to pass around code fragments. Over the years I've put a lot of weird little things on mine, and they form a little history of my hard thinking. Thoughts I couldn't have had (or couldn't have finished) without Python.

---

## Via Julia Galef: comparing Spock's predictive skill to a coin flip

The online forecasting community have a way of scoring how calibrated someone is, the Brier score. The *Enterprise* would do better to consult a random process like a coin flip (and in fact they do ignore him most of the time). This matters because Spock is one of the main exemplars for rational thought, and he is a dingbat, which might explain some people's attitude towards explicit rationality.

A natural objection: the episodes we see are not a representative sample of events; they are “selected” to be as dramatic and telegenic as they can be, which

means that of course Spock predicts wrong! But by the exact same token, these are the events it is most important for Spock to predict well, and he does terribly, so the expected value of listening to him is still bad even if he gets everything else right.

(My preferred explanation is that Vulcans are in fact stupid, they just talk like smart people.)

### **An optimal solution to a friend's data adventure game.**

Fun! Only interesting because it includes expected value and risk-sensitive optimisation solutions. EV alone would not have done well at this task; I gave it only 20% of the budget.

The death sensitive bit is here: just variance (z-score) as a danger score.

### **Kelly for maximum house insurance cost**

The Kelly criterion is an interesting piece of abstract nonsense that tells you how much you should bet / pay for insurance, given how much money you have, if losing all your money means death / infinite loss. It takes a bit of work to compute it, but it gives surprisingly intuitive answers, and it beats the hell out of EV when there are big downside risks.

### **Reported vs real (excess) COVID deaths, 2020**

One of the 700 bitter stupid information-free arguments people had about COVID early on was the “infection-fatality ratio” - was it 0.1% like flu? 3% like SARS? I did this script to show that the ascertainment rate (the % of cases you catch in your system) isn’t constant, in order to justify a modelling choice, as part of me losing a year pretending (quite well) to be an epidemiologist.

### **Converting between the effect sizes used in social sciences**

A quiet improvement in psychology over the last 15 years: people started to foreground the actual size of the effects they studied, and to use nice standard metrics for em. (Part of my inexplicable procrastination on my PhD quals was dissing 300 psychology papers.)

### **A really nice way to juggle probabilities: dict keys**

Solutions to Chapter 2 from the mighty mighty ITILA.

Also the classic minimal case of Bayesian updates, coins and binomials.

### **Google character recognition**

I recently scanned in my teenage handwritten notebooks, and tried out the GCP OCR. It’s free up to a few thousand requests. Accuracy is surprisingly not great;

corrections take about 15 minutes per small book. If my time was more valuable / if it wasn't a joy to look at these dumb notes again I might try fine-tuning a Transformer on my handwriting.

### **Helping a friend maximise his alcohol purchases**

This was part of my trying to sell him on the value of programming, god save him.

#### **Tax check**

Boring script to see how much teaching I can do without incurring tax.

#### **Queue as Poisson process.**

Incredibly involved exercise from BDA Chapter 1. Not conceptually difficult, just very fiddly.

#### **Check power of 3 in near-constant time**

$$3^x \leq 2^{63} - 1$$

take  $\log_3$

i.e.

$$x = \log_3(2^{63} - 1)$$

We know

$$3^x \% 3^y = 0$$

for  $y < x$

So for a near-constant time check, just ensure that you make

$$3^x$$

definitionally larger than

$$3^y$$

. e.g. set it to the MAX int of your system, and then take the mod.

# Songs for the Extremely Online

Gavin

2022-01-22

```
{% assign nbt = "https://www.youtube.com/watch?v=wfBdL32L3Z8" %} {%  
assign burnham = "https://www.youtube.com/watch?v=lBQalWCH-hw" %}  
{% assign aero = "https://www.science.org/doi/10.1126/science.abd9149" %} {%  
assign peter = "https://www.youtube.com/watch?v=wLoG9zBvvLQ" %} {% as-  
sign forgues = "https://journals.sagepub.com/doi/pdf/10.1177/1476127012452820"  
%}
```

Could I interest you in everything, all of the time? A little bit of everything, all of the time? Apathy's a tragedy - and boredom is a crime! Anything and everything: all of the time.

– Bo Burnham as The Internet

Because the Internet, mistakes are forever But if we fuck up this journey, at least we're together... No one's ever been this lost I just get the information, retweet or say it sucks

– Childish Gambino

What does music tell us about the world? Almost nothing.

What does music tell us about the prevailing view of the world? Something: for most people it's more ideologically powerful than books.

The cool music of the noughties has a clear worldview. Too clear, if anything, but then the evils seemed clear. Anti-war, anti-surveillance, atheistic, pirate parties and information wants to be free. Freedom as a bipartisan or anyway double concept. The mood is agentic: "sure our foes have the power now, but if we *do* something, if we just have enough journalists and scientists and NGOs, we'll win." Moral clarity, even if naive, even if loud-mouthed.

What zeitgeist is embodied now?

Malaise. The songs I cherrypicked for this post are Not Okay, or whatever. They think this is to do with the internet, or the poisoned media environment, or the poisoned air, or whatever. These musicians are Extremely Online outliers and cannot stand in for the young, psychoanalysis of them does not generalise, does not characterise the default ideology of the whole generation, or whatever.

But that said you can hear the apocalyptic apathy, the apathetic apocalypsim.  
(All about the end.) Depressive flagellants. (“*My tuition's paid by blood, I might deserve your fate or worse.*”)

```
<h3>Incomplete cultural analysis of the last decade</h3>
<div>
  <b>What's wrong with what you think's wrong with the world?</b><br>
  <!-- -->
  * Anti-tech<br>
  * feeling of unreality.<br>
  * Post-truth. Misinformation. Russiagate paranoia, first-order programmable psychology.<br>
  * Climate change.<br>
  * Anxiety, weakness, fatigue, uselessness, guilt. To some extent this is just increased
    What does it do, to stare constantly at your own fallibility and hopelessness? To attack
    Anti-tech without the consolation of primitivism or traditionalism. Unable to log off be
    Somehow the top priority of the default worldview, racism, doesn't come up in these song
  <b>What changed in 10 years?</b><br><br>
  • Rate of interacting with anon strangers
  • The left switched to personalising social problems. (Strictly speaking the
    theories are all still structural, but the praxis is radically individualist and
    moralistic.) Fuck Peterson: Marx is turning in his grave.
  • Podcasts: The return of radio! The triumph of the crowd! Zines that
    people actually care about!
  • Online dating dominates, sexting
  • The death and rebirth of blogging
```

---

#### *because the internet* by Childish Gambino

Man made the web, you don't need a name  
Man made of faults, I ain't  
too ashamed... Every thought I had, I put it in a box  
Everybody see it, just before the cops Andrew Auernheimer pulling on her weave,  
it's that Andrew Auernheimer Texts from people I never met, doors  
left open (Who is this? Don't do it, where are you? Who is this?  
Who is this?) I don't know who I am anymore

This album is nearly 10 years old, and it certainly feels different, dated (e.g. name-dropping “iphone”). But CG shares the essential queasiness of the rest of this internet music: he was clearly at the vanguard of the extremely online.

The mood is ultra anxious, decadent, but unlike others he manages to convey some wonder and love of info and tech.

### ***Inside* by Bo Burnham**

A covid album more than an internet album. But the covid years were the most intense internet years ever.

He is always skilled, sometimes beautiful - but often limits himself to novelty songs. (What is a novelty? Something which works only once, regardless of the quality of that once.)

Mommy let you use her iPad; you were barely two And it did all the things we designed it to do Now, look at you! Oh, look at you! You, you! Unstoppable, watchable...

*White Woman's Instagram* is a few things. It sneers. It's crabbed: seeing positivity makes him want to drag it down. It hides its basic negativity behind political piety: without the 'white' qualifier to validate it, it would have caused him trouble, "joyshaming" or whatever.

But then it drops all that and shows the woman mourning her dead parents and I defy you to keep smirking.

Her favorite photo of her mom The caption says, "I can't believe it It's been a decade since you've been gone... It's got a little better but it's still hard Mama, I got a job I love and my own apartment Mama, I got a boyfriend and I'm crazy about him Your little girl didn't do too bad Mama, I love you, give a hug and kiss to dad"

(Although I'm not actually sure that cruelty was the base level intention, rather than cruelty-baiting. *If in fact* he was pretending to join in with our contempt - to get us to lower our guard and so feel ashamed when those dead parents make us remember that this is a [fictional? no matter] person and that taste-as-in-cynicism is a small thing - then this is greatness. But I fear not.)

What torments Burnham? Lockdown, clearly. Modern political guilt, but only somewhat. Something organic under it all.

*Problematic* is perfectly poised between sincerity and irony ("Or that I'd live to regret it"). The great moment is when the second half apologises for the first half of the song, an acute look at a very common online phenom, where someone apologises and then has to apologise after the initial apology gets savaged.

I want to show you how I'm growing as a person, but first I feel I must address the lyrics from the previous verse I tried to hide behind my childhood and that's not okay My actions are my own, I won't explain them away I've done a lot of self-reflecting Since I started singing this song I was totally wrong when I said it Father, please forgive me for I did not realize what I did (Or that I'd live to regret it) The times are changing and I'm getting old Are you gonna hold me accountable? My bed is empty, and I'm getting cold Isn't anybody gonna hold me accountable?

But the song also distances him from his stupid, brilliant early work. Honour him for not allowing this to *utterly* crush him, for still telling some jokes, even if they're sad and reject comedy.

Stunning 8K-resolution meditation app In honor of the revolution,  
it's half-off at the Gap Deadpool, self-awareness, loving parents,  
harmless fun The backlash to the backlash to the thing that's just  
begun There it is again, that funny feeling... Twenty-thousand years  
of this, seven more to go... Full agoraphobic, losing focus, cover  
blown A book on getting better hand-delivered by a drone Total  
disassociation, fully out your mind Googling derealization, hating  
what you find That unapparent summer air in early fall The quiet  
comprehending of the ending of it all There it is again, that funny  
feeling That funny feeling Hey, what can you say? We were overdue  
But it'll be over soon

I expect this to age much better than Childish Gambino.

(BTW the visual design of the show is more than half of the artistry, so watch rather than listen.)

## ***Moral Panic by Nothing But Thieves***

I can't pin these guys down. 'Moral Panic' is a great title for satire - but they're too earnest and don't seem able to rise above their moment. e.g. There's a random dig at MAGAs. (They're English.)

More like Pendulum than Mclusky. Still, one great song:

I fucking hate the internet The fame suckers in their block-long cars  
Five star hotel (i don't feel well) I think I'll cancel the honeymoon...  
I could use some healing soon Before I lose all feeling soon...

We're shutting down the internet... I got some pills but not some  
help make my clicks spike Why don't we hit the minima?

## ***public void by Penelope Scott***

Remarkable. Scott is both hyperintellectual and anti-intellectual, a radical with no theory and no praxis. (She has much the same revulsion towards formal microeconomics as I once did. The title is a Java joke.)

Atomistic rational behavior / Invisible hand savior Fucking up your  
definitions even though it's life or death Who fucking told you you  
were selfish Or even self-interested Don't you think it matters when  
we're with our friends, the best

The obvious standout is *Raet*, an elegy of a Musk fan who woke up, a post-rationalist anthem. But *Moonsickness* is her lyrical triumph, or whatever the opposite of triumph is: exultant despair, proud sickness.

I've got one hundred hours to rearrange the stars  
And I'm the worst  
mistake that God has ever made  
You seem to integrate so fucking well  
But I make lemons out of lemonade...  
If you had children now you  
think You might just put them down  
None of us belong  
Everything I  
do is wrong  
And fuck I'm not a Marxist I'm not a fucking democrat  
Because of all this bullshit I'm not anything at all  
All I wanted was a  
framework  
None of them can live here  
There's nothing to believe in  
and there won't be til we fall  
And it's not all you man  
You were just  
a kid once  
God I'm such a fuck up...  
I've got one-hundred hours to  
rearrange the stars  
And I'm the worst mistake your God has ever  
made  
I can't get the numbers right  
I can't fucking count because not  
one goddamned thing is in its place

Elsewhere she addresses the bizarre feminine love of true crime podcasts. (Google Trend for “true crime” against US violent crime rate...)

### ***Solid State* by Jonathan Coulton**

Rare entry from the prior generation, someone with a reference for what things were like before. Cancel culture exists here. But actually it's only half current malaise and half an oddly detailed picture of a post-human malaise. (“*I lit up the sea, pulled down the stars for you*”) “Sunshine” is a remarkable portrait of a Disneyland without children, sung by the final unemulated humans. Ray Kurzweil comes up in two different songs. You may take my mortality, you may take my toil, but you shall never take my misery.

Small mistake: “a terrible crime”. It’s better than solid state. It’s all messed up, it’s better that way Everyone you know, crooked little numbers game Everywhere you go, it’s all the same Watch them rise and fall Human after all (take care of other) Used to be, the world was too far away Used to be, the stars didn’t have much to say

(Is the eponymous ‘solid state’ death? No, it’s larger: it’s the state of nonexistence: when your legacy, all evidence of your past and any continuing effects of your actions cease.)

The echo of a choice, the static that you leave behind / Is better than solid state.

*Brave* is a portrait of the keyboard warrior, mostly the incel kind (though the right has no monopoly on resentment). It’s not totally devoid of empathy.

Slack-jawed sheeple with their eyes closed There’s too many of you,  
more than I can save When I torch the place, cover up my face,

That will make me brave. Filling in the shapes of shadows in my cave... You speak and presto-changeo, now I'm the bad one. My heart hardening, counting up the lonely nights, all the little slights I'm taking to my grave.

#### *Pictures of Cats*

All of the pieces and none of the places they go So I am looking at pictures of cats. Too close, so I'm pretending I'm far far away. Not now; I didn't want to be useless today. Try me tomorrow, today has been laying me low.

#### *Don't Feed The Trolls*

The best depiction of the chilling effect.

The other artists mostly ignore surveillance, one of the defining evils from before. They feel they have larger problems

Dance like they're watching you, because they are watching you.  
And when the bright lights find you, don't let your heart get lazy  
Don't read the comments and don't feed the trolls... Appreciate the outrage, I did the best I could I thought about your thinkpiece, I don't think it's any good I just checked my privilege, and it looks fine...  
And when the bright lights find you, bro come on bro don't taze me  
Don't read the comments and don't feed the trolls. Don't read the comments and don't feed the trolls. Don't read the comments and don't feed the trolls.

To listen to them, you'd think the internet was a curse.

I'm reminded of the popular view of Twitter (and Reddit), as a sea of toxin which needs to be tightly controlled. Sorry to tell you that the toxin is coming from inside the house.

None succeed in capturing the internet's aesthetics and logics as well as Dril. But then nothing does.

```
<h3>Cherrypicking and lemonpicking</h3>
<div>
    Shockingly good writing from a business professor, <a href="{{forgues}}>Forgues</a>:
    <!-- -->
    <blockquote>
        The problem is, extreme cases are rare. Rare enough not to show up in our random samples
    </blockquote>
    <!-- -->
    So too with all cultural criticism, like the above.
</div>
```

### **See also**

- Tyler on music as vehicle for ideas - formerly.

# Cracking cultures

Gavin

2021-10-25

{% include cultural/links.md %}

There is so much to understand, there are so many things to like - and all of us understand and like very little of it. 2 If you're young, your preferences maybe define you - but you could be defined by something else. If you're any age, they determine who you spend your time with - but you could spend time with almost anyone.

It's easy to forget how inexhaustible the natural and human world is. This year I've been running classes on *the art of getting into things*. I collected all the interests and subcultures my students are into (or actively not into):

{% include cultural/list.md %}

You can attempt to get into any such system of meaning. What's a word for those? "Culture" - but they can be *much* smaller than the national units we usually mean by "culture". So here take "a culture" to mean a subculture, an idiom, a scene, a style, a genre, a field, a medium, a view.

Claim: every human activity, and every group of humans larger than one, forms a culture. It's often intentionally hard for outsiders to understand. Reality has a lot of detail, and humans, good humans, paint this detail with meaning and distinctions. Cracking these codes is the most important skill which is barely taught anywhere. 5

Every year I try to get into something major. Hacker lore (2014); "modern classical" music (2020); Chinese poetry (2011); Analytic philosophy (2008); economic rationality (2013); dank memes (2017); singing in public (2016); teaching (2020); this year, comic books. 3

{% include cultural/cowen.md %}

## Why?

### 1. Fun! Access more of the value in the world.

It's more than just liking more cool stuff. It's about treating your own taste and interest as an object in question, an object which could be worked on. About

treating outgroups as puzzles rather than threats or weirdos.

## 2. Social life

What you like determines who you spend your time with. It often determines your life partner. Businesses and academic fields are famously culturally ornate and hard to crack.

You've probably had the experience of being at a party and realising that the stranger next to you shares your love of Japanese noise rock, or loose-leaf tea, or Afro-futurism, or Adult Swim. You've probably had some amazing conversations as a result. This tells us we can deeply interact with 10x more strangers.

## 3. Understand people!

- You break off a piece of the giant impossible concept of human culture overall.
- Subcultures do a large amount of all new and interesting work. (This is almost true by definition in art. But also startups: to replace a huge corporate incumbent, you have to have a different angle, and often they are outsiders.)
- *Mental flexibility.* One of the evils of ageing is bewilderment: feeling that the world is bizarre and unmanageable, that you can't interact with the young, that you are relegated. Active effort and mastery of cultures should prevent this.

{% include cultural/techpay.html %}

### Why aesthetics?

Above, I made grand claims about large portions of all human activities being available to crack. 8 So why am I talking about comics?

Aesthetics is a great place to start because it's so cheap and the experiments are so quick. It's also surprisingly impactful, socially powerful. 4 And also because once you stop seeing your taste as immutable (or god forbid correct) you can pursue all of the rest of the world.

I really think there's a general skill here - that understanding punk deeply really does increase my ability to understand Tanzanian culture, let alone prog and disco and post-punk and dub and thrash.

## A spectrum

Three ways of relating to a genre, a medium, an art, a school of thought, a field:

Love: to find value in ordinary examples.

Open to: to see the value of the best examples.

Not open to: to struggle to see the value of even the best.

I claim that 0% and 100% are basically never correct. Most things fall into 30-40%. We want to move from 20% to 60% on most things. (It would be very distracting to love everything.)

---

Pick something you're not open to. Ask:

- Why don't I like it?
  - Do I not like the people who like it?
  - Do my friends dislike it?
  - Does it offend me? Is it ugly? Low class? Pretentious?
  - What is it trying to do?
  - How would I have to change to get it?
- 

## Actionable bits of a cultural code

To make something interesting, just look at it a long time.

— Gustave Flaubert

### 1. Canon.

You need to start with the greatest (or the most accessible greats) so that you can remove one source of uncertainty and solve for the remaining unknown: your stomach for it. Some cultures revolt against the idea of a canon, and but all of them have secret shibboleth canons behind the listicle canons.

Finding critics you can trust helps, because their activity consists in taking the unwritten and writing it down. (Obviously they never fully succeed.) Outside literature, academics are often surprisingly poor critics.

Once you know the canon, you can get the *allusions*, and you can understand the principal components, the ways instances are supposed to vary.

### 2. Jargon, conventions, techniques

Too specific to say much about here. Critics again, or else a hard act of empiricism.

### 3. Material conditions

Say you try it. Say you pick the top 10 all-time whatevers. But you bounce off - it all seems so contrived / so hostile to its audience / so trivial / so pretentious. What to do?

One powerful trick is to study what Marxists call the material conditions. 1 For our purposes this is not a grand reduction of the ideal to the economic, it's just 1) how capital-intensive it is, 2) the demographics of the creators and audience, 3) the tempo and complexity of production (weekly for manga, a month for a serious poem). Then: *how do they do it?* Those timelapse videos of someone painting or carving are ideal.

I watched every Kubrick film and didn't really see the fuss. Then I read up on him, and learned that e.g. he had thousands upon thousands of doorways in London photographed while location scouting for *Eyes Wide Shut*. It's not that obsessiveness means quality, that inputs mean output. But it means *meaning*. As I rewatch him, I have good reason to consider many parts of the production as meaningful, and in fact I like him far more on the second runthrough.

Only once you know what's good, what the axes are, and how it's made can you understand originality, deviance, substyles, and your own sense of the greatness.

---

## Procedure

1. Find a critic you can trust. Friends are best. [2 weeks]
  2. Where is the quality? What is it *trying* to do? [2 weeks]
  3. What are the material conditions? [2 weeks]
  4. If you really can't see any value: what's sociologically remarkable about it?
  5. When do I just accept that I am not capable of liking this? 6
- 

## Examples

```
{% include cultural/poetry.md %} {% include cultural/punk.md %} {% include cultural/ea.md %} {% include cultural/an.md %}
```

---

What will you like? What will you understand?

---

## See also

- Am I advocating being a mop?

- Seeking Sense
- Comfort Zone Expansion
- Zohar Atkins trying to explain dense and repugnant thinkers in plain and alluring terms.
- Callard on aspiring, Callard's own idioms
- Logan on poetry
- On self-invention more generally
- On the cultural code of corporations

{% include cultural/limits.md %} {% include cultural/act.html %} {% include cultural/chain.html %}

{% include cultural/foots.html %} {% include lazyload.html %}

# Trying to do some good, a history

Gavin

2021-02-22

Political brained. Chomsky, Klein, Greer, Bakan, Pilger. Money is crass. Wealth is strong evidence of exploitation. Killing the planet.

Refuse to apply to Oxbridge despite being invited.

2007: Apply for Biology, Music, Japanese, English, Philosophy. Get in to all of em.

Sep 2008. Arrive at uni to do music. Switch to economics.

Sep 2011. Vegan

May 2012. Graduate. Deeply aware that I need to get technical.  
<https://afterallitcouldbeworse.blogspot.com/2012/05/sentimental-graduate-22-seeks.html>

Summer 2012. Volunteer on a gender equality project in Tanzania.

Oct 2012. Start a part-time maths degree.

Jan 2013 80k advice: Jess W Reject the idea of earning to give.

Rejected for various economics Master's.

Learn to code.

Feb 2015. GWWC pledge

July 2016. Blag my way into a data science job.

Nov 2016. EAGxOxford. Meet Nintil.

Get EA friends hired.

April 2018. AI Safety Camp.

October 2018. x risk

Dec 2018. First EA Forum post.

Feb 2021. First AI safety paper.

July 2021. Start an EA org.

Sep 2021. Teach at ESPR.

# ‘Terra Ignota’: the ecstasy of uncertainty

Gavin

2021-11-29

{% include palmer/links.md %}

Palmer’s series suggests [that] science fiction should not be viewed as just another literary genre, but as the genre where Enlightenment—the hopes for radical human self-improvement, the dream that we might collectively control our own fate as a species, the determination to transcend our own limitations—takes refuge in an anti-Utopian age that seems determined to deflate any such ambitions... Its ambitions and achievements far surpass... the limited imagination of fictions that confine themselves to representing everyday life. More than philosophy or political theory, science fiction is the genre through which our age joins the Great Conversation.

— Lee Konstantinou

Regarding “Terra Ignota”, a series of novels by Ada Palmer:

The series is a lot of things. It is the most sustained fictional portrait of Archipelago and polystates, one of the few utopias I would maybe like to live in. Palmer starts in an Enlightenment utopia (post-war, post-nationalism, post-scarcity, post-gender, post-theocracy, post-fideism, post-meat, post-capital-punishment, post-nuclear-family, general justice via universal voluntary surveillance) and then shows what the tensions will do to any system that has to handle humans as we are.

The worldview diversity is probably the greatest thing about it. I’ve read twenty-author anthologies with less variance in values than this. Speaker’s Corner and SSC comments have nothing on Palmer. You think I’m being bien-pensant right now, praising diversity - but there are fascists in it! Sex-murder teens! The Worst Fan In The World! Rapist priestesses! All

About half of readers find the prose unbearably clotted and affected. (If you’ve read books from more than two hundred years ago you’ll have some immunity.) I loved the many didactic discourses - e.g. de Sade’s Christian name being a plot point, sections written in speculative future Latin - but I think most readers will not love them. You’ll have to be fine with long fourth-wall violations, long passages in macaronic Latin, hallucinated philosophers reacting to C25th scenes by expositing their extrapolated view of the 25th Century, allusions that yell

‘REMEMBER ME??’ in your face (Hobbestown, the anarchist commune). I found the narrator’s madness engaging but it does divert every chapter a bit.

(Meme: “in the grim darkness of the C25th, mankind has divided into its elemental archetypes: jock, fash, hufflepuff, freud, stemlord, landlord, libertarian, person with a country of origin instead of a personality, and ‘meh’”. This is no critique of Palmer when we remember that all such groupings will arise through partially random historical contingencies: the resulting categories don’t need to make sense and probably won’t.)

The books depict superpowers, *even if we ignore the 2 or 3 supernatural beings*. The Mardis, the Censors, and the set-sets have ridiculous amounts of predictive power using Weird Data Science, predicting the timing of world events 20 or 30 years out. The Brillists have this power, plus mind reading, and bizarro mind control, and arbitrary hacking power. These are *so* much more powerful than the tame AIs and giant mechas of the Utopians. But the plot is unchanged by them until the last book, at which point they are easily subverted for confusing reasons.

“Worldbuilding” is often a red flag. It predicts an author who cares more about their lore than their characters or plot, who is going to fail to make you *care* that the legal system or the conlang or the magic system is consistent. Palmer is the queen of worldbuilding, *and yet she gets over it*: her characters somehow nevertheless rule the series. It is quite obvious that large amounts of her notes did not make it into the 2000 pages of this series.

I could see you, across the sky, the crowded sea, a thousand black and winged shapes for every tardy, well-meant [dove]. But humans began digging a canal across the Gulf of Corinth more than three thousand years ago and finished it in 1893. It’s worth trying things again. Apollo Guardian of Strangers knows that it’s worth trying things again. Especially for [peace].

## **Book 1: *Too Like the Lightning***

I choked a little at the constant coincidences, and at the enslaved protagonist meeting literally every elite in the world in the space of two days. (“Providence” innit.)

{% include palmer/tltl.md %}

## **Book 2: *Seven Surrenders***

The sunny, war-free Hive system gets subverted multiple times. The Cousin democracy is fake. The Masons get exposed. But every Hive is governed at the whim of Madame and her captive orgy. Missed the first time: The Madame conspiracy are as bad as you’d expect, silently squeezing the pluralism and democracy out of the world

Perry has been a midlevel member of this establishment for six years now. No one could advance so far in politics without some help from here.

## **Book 3: The Will to Battle**

{% include palmer/wtb.md %}

## **Book 4: Perhaps the Stars**

{% include palmer/praps.md %}

### **See also**

- Robnost trying to understand why the bad bits are there
- dril

{% include palmer/foots.html %}

# Favourite maths tools

Gavin

2021-11-02

```
{% assign approach = "https://approach0.xyz/search/" %} {% assign open = "http://www.openproblemgarden.org/" %} {% assign i = "http://oeis.org/" %} {% assign charton = "https://www.quantamagazine.org/symbolic-mathematics-finally-yields-to-neural-networks-20200520/" %} {% assign mit = "https://ocw.mit.edu/courses/find-by-topic/#cat=mathematics" %} {% assign box = "https://github.com/mentat-collective/mathbox2/blob/master/docs/intro.md" %} {% assign man = "https://github.com/3b1b/manim" %} {% assign g = "https://nathancarter.github.io/group-explorer/GroupExplorer.html" %} {% assign gr = "https://www.graphclasses.org/classes/gc_72.html" %} {% assign cat = "https://ncatlab.org/nlab/page_categories" %} {% assign mo = "https://mathoverflow.net/" %} {% assign des = "https://www.desmos.com/calculator/tdhxorkxgb" %} {% assign w = "https://www.wolframalpha.com/" %} {% assign py = "https://www.sympy.org/en/index.html" %} {% assign max = "https://maxima.sourceforge.io/" %}
```

## Writing

I'm open to the idea that pen n paper are superior to typing, for *learning*.

- But I stick everything in Roam. Nested lists, LaTeX, PDFs, images and videos, all in one doc.

```
<h3>e.g.</h3>
<div>
  <br>
</div>
```

- I am no longer sure if it's worth learning LaTeX, but if you want to or have to, for the love of god use Overleaf. Saves hours a year on install headaches.
- Years and years in, I still sometimes encounter symbols I don't know the name for. Detexify has saved me many hours
- MathPix is an AI maths recogniser, so you can go from e.g. handwritten notes to Latex. Very good accuracy, though the formatting often needs cleaning up.

## Computer algebra (“solve this for me”)

- WolframAlpha. Obvs! Pro is worthwhile for absolute beginners, gives you infinite stepped examples.
- Maxima. If you need something heavier and free. I see no point in Mathematica or SAGE or even Octave.
- SymPy. Satisfying but not productive. Only if you’re 10x stronger at programming than maths, as I was.
- I’m surprised there’s no product of this yet (neural net beats Mathematica on univariable integration)

## References

- I’ve probably spent more time on MathOverflow than I have in textbooks.
- Search engine for formulae. Wolfram does this a bit too.
- Catalogue of groups
- Catalogue of graphs
- Catalogue of categories
- I’ve never actually used OEIS but some people seem to get real life use out of it.
- Currently open problems ranked by importance

## Viz

- draw.io is way better than it looks. You can easily make publication-quality vector graphics. The good bits of Powerpoint.
- Loads of graphing tools at Desmos (e.g. teaching linear programming in 2 variables)
- Geogebra looks fine too
- Never used either, but for generating beautiful video with a lot of work you now have MathBox and Manim

## Challenges

- Project Euler is the second best way to learn a new programming language (once you know one deeply). Some people do way better at the computational than the analytic, and this is one bridge you can take.

## Courses

I have only ever liked and finished one MOOC (and in general I dislike learning from video).

- But people seem to love OpenCourseWare.

**See also**

- Ranking lots of linear algebra material

# Is whey ok?

Gavin

2021-11-22

I'm vegan. But I'm a confusing sort. It's not putting animal products in your mouth that does the damage, it's paying for them.

If I had a stronger stomach, this means there's nothing wrong with eating discarded meat.

<https://www.jefftk.com/p/how-bad-is-dairy>

"1 cup of low-fat milk contains 8.53 grams of protein, of which 18% is whey protein, implying that 1 cup of milk contains ~1.54g whey protein. So to get 25g of whey protein (the recommended dose for a standard Power Smoothie) would require about 16 cups of milk (25/1.54), or about 2000 calories at 124 calories per cup. That means that if a cow produces 17,640,000 calories' worth of milk over the course of its life, it produces the equivalent of ~8700 scoops of whey protein powder, assuming that the protein is isolated perfectly. This would imply that having one Power Smoothie a day for 24 years would account for one cow."

"While this estimate does assume perfect isolation of protein, it's also worth noting that whey is a by-product of cheese production, i.e., the other parts of the milk are used too and may actually be in more demand than the parts used to make whey protein powder. Thus, I think this estimate is more likely to overstate than underestimate the impact of whey protein powder consumption, and overall I'd be surprised if anyone could consume enough Power Smoothie in their lifetime to account for 2 cows (at a few Power Smoothies a week currently, I doubt that I will account for 1). Contrast this with chicken, in which it might only take 3-4 chicken-heavy meals to account for one chicken's life."

Lewis Bolland: "You should change from comparing "calories per life," which is hugely skewed by lifespans — 4-6 years for a dairy cow vs. 35-55 days for a broiler chicken — to "calories per day." (Assuming you agree with me that suffering, not lives lost, is what matters.) For dairy cows, that would be about 15,000 calories/day (avg. 6.5 gallons of milk/cow/day \* 8.5lbs per gallon \* 272 calories per lb of whole milk — these are conservative numbers; some stats put gallons/day higher). (3) Given that whey seems to be either as in demand, or less in demand than, other parts of the milk, it seems to make sense to just count the calories of the whey (assuming the other calories in the milk will

be used for cheese etc.). (Note these numbers assume that the whey protein concentrate provides the same average value per calorie to producers as other components of milk. I haven't found great numbers breaking down the relative value of cheese and whey sales to cheese producers, who produce whey as a byproduct. But within the whey market, the prices for "Dry Whey Central" (used for humans) and "Dry Whey Animal Feed" are surprisingly similar. This suggests that human use whey protein is not disproportionately profitable to dairy producers, though that could change if demand surged since whey protein concentrate supply is limited by the inefficient process used to produce it.) The power smoothie requires 25g of whey protein, which is 100 calories (just from the protein — presumably they soup it up with fat and sugar from non-dairy sources, so not counting those). That implies each power smoothie contributes to just 1/150th of a cow's daily milk output. I.e. if you have a power smoothie every day for a year, you'll only be responsible for 2.4 days of a cow's life on the farm, and if you have a smoothie every day for the next 30 years, you'll only be responsible for 70 days of a cow's life. (4) If the contribution to demand is really this low, I think the other aspects of the dairy industry are pretty irrelevant. My understanding is that the average dairy cow is kept for about 3.6 births, and the replacement rate is just over 1 because mortality is low, so perhaps 2.5 calves are killed near birth per cow over its 4-6 year commercial lifespan. So if you drank power smoothies every day for 30 years, you'd only be responsible for about 1/10th of a calf getting killed. Hope this is all good news!"

# Unthinking meat

Gavin

2021-01-31

“You’re saying they have an exquisitely sensitive and accurate sensory apparatus, and an unbounded memory capacity, and fully general problem-solving faculties?”

“Well, sort of:

“When they’re not focussing, which is 95% of the time, they can’t really be said to be intelligent at all. Much of what they say and do is hollow reflex motion.

“They also fill most of their bandwidth up with information which is worthless at best and usually actively misleading. They find fabrications more convincing than data. They rarely do what they think is most important.

“They also keep their current sense data, memories, moral evaluations, aesthetic evaluations, and political evaluations - their lust, fear, and avarice - all in the same chamber. This makes them confuse fact with value, rights with wishes, and desire with everything.

“Most of their lives are spent on coalition maintenance, social grooming, and monitoring and enforcing hierarchy.

“They have no access to much of the most action-relevant parts of their processor, which has developed backdoors to systematically delude the narrator about the system’s goals and motives. They are in effect incapable of honesty.

“While the processor is capable of running formal logic, very very slowly, in practice they use a series of appallingly non-Bayesian evolutionary algorithms to do almost all of their reasoning, including about the central concerns of their lives, mates, careers, and finance.”

“... not what you’d call a threat then.”

“Well, not to us.”

## See also

- The melancholy of pareidolia
- Why is quality rare?
- Pieties
- Where does reason end?

- Heuristics, cognitive miser, attribute substitution.
- Simler, Elephant in the Brain
- Constantin, Humans Who Are Not Concentrating Are Not General Intelligences
- Crichton, Gell-Mann Amnesia
- Taleb, Against News

# Failed epistemic hygiene measures

Gavin

2021-12-02

- The social role of the expert
- The social role of the fact-checker (whitelisting people)
- Meta analysis
- Debiasing
- Social media NLP censorship (whitelisting opinions)

A public health for the information ecosystem. Some public health bodies and mouthpieces acted clownishly, and bristled at any challenge to their authority.

## Pre-failure

- Prediction markets
- FullFact auto annotation
- StackExchange

They failed to solve the whole problem. But why expect any one thing to do that?

# Ranking linear algebra courses

Gavin

2021-12-02

Video? Hardcopy? Exercises? Visualisations? Interactive?

All the Mathematics chap 1

Axler Beezer Linear Algebra Done Wrong \* 6 Coding the Matrix' by Philip N. Klein - immersive linear algebra - Graphical Linear Algebra <https://aiprobook.com/numerical-linear-algebra-for-programmers/> <https://hefferon.net/linealgebra/>

<https://brilliant.org/courses/linear-algebra/> <https://app.bolster.academy/courses/chapters/en/6> <https://minireference.com/>

## See also

- HBPMS
- Abhishek Roy on hundreds of books

# Christmas feel

Gavin

2021-12-10

```
{% assign g = "https://www.theguardian.com/music/2019/dec/23/observer-readers-alternative-christmas-playlist" %}
```

## 1. Ruminate

wound-licking

## 2. Hibernate

No obligations. The failing of strength, or husbanding.

Whisper.

## 3. Innocence, or regression to innocence

## 4. Family

## 5. Cynic, degenerate, based, imp of the perverse

I am amazed by how people think they're clever for thinking of talking about dark things at Christmas, or for noting that commerce and culture often go together in unclear causal directions.

## After cynicism

You'll never be innocent again. In some ways the iron has entered you. In some ways you feel less than you used to. But you don't have to go gently. You don't have to pretend that this is strength, or sophistication, and that actually you haven't lost anything.

## 6. The New Possibility

I am an almost total void of spiritual feeling. I struggle to regard anything at all as sacred. Music gives me an echo.

John Fahey, a believer, called his xmas album *The New Possibility*: the hope borne of the turning point, not of one year, but of all of history.

‘Sound, Sound Your Instruments of Joy’ strikes me as one of the finest things I’ve ever heard.

‘Sister Winter’ is hibernation followed by jubilation.

Russell

Gavin

2017-09-01

# Ramsey

Gavin

2022-02-18

(A ramble occasioned by Cheryl Misak's biography, but not a review.)

```
{% assign iar = "https://en.wikipedia.org/wiki/I._A._Richards" %} {%  
assign br = "https://en.wikipedia.org/wiki/R._B._Braithwaite" %} {%  
assign k = "https://en.wikipedia.org/wiki/John_Maynard_Keynes" %}  
{% assign boi = "https://www.jstor.org/stable/1905256" %} {% assign  
dm = "https://www.jstor.org/stable/1910538" %} {% assign ben =  
"https://eprints.whiterose.ac.uk/132301/1/bentham's_binary_form_of_maximizing_utilitarianism.pdf"  
%}
```

[Ramsey, aged 17] turned to [CK] Ogden and said: 'Do you know, I've been thinking I ought to learn German. How do you learn German?' Ogden leaped up instantly, rushed to the shelf, got him a very thorough German grammar — and an Anglo-German dictionary — and then hunted on the shelves and found a very abstruse work in German — Mach's Die Analyse der Empfindungen — and said: 'You're obviously interested in this, and all you do is to read the book. Use the grammar and use the dictionary and come and tell us what you think.' Believe it or not, within ten days, Frank was back saying that Mach had misstated this and that he ought to have developed that argument more fully, it wasn't satisfactory. He'd learned to read German — not to speak it, but to read it — in almost hardly over a week.

– Richards 1

"A Mathematical Theory of Saving" . . . is, I think, one of the most remarkable contributions to mathematical economics ever made, both in respect of the intrinsic importance and difficulty of its subject, the power and elegance of the technical methods employed, and the clear purity of illumination with which the writer's mind is felt by the reader to play about its subject.

– Keynes

There was something a bit abnormal about Frank. He was so huge in body and in mind, so much bigger and better than the rest

of us, that I suspected that... his cells might have double the number of chromosomes as those of ordinary men... While still an undergraduate Frank had attacked Keynes on the subject of his theory of probability and had shaken him to the core. But this precocious intelligence was combined with a childlike innocence... The result was very curious. When I brought to his notice some ordinary tale of petty self-seeking, self-deception or malice, Frank was at first astounded. Such things did not seem possible to him up there in the heights. Then he would realise the full implications and humour of folly and silliness, and the self-defeating nature of selfishness and spitefulness, and God-like, his great innocent face would become wreathed in smiles and then he would chuckle. And his chuckle was the chuckling of a god.

– Braithwaite

Franz Berto calls Ramsey “the counterfactual greatest philosopher of the 20th Century” (i.e. he would have been that, if he hadn’t died at 26). And he’s one of my favourite people, a model personality as well as a titanic brain. He was just a big jolly bastard. It is right to call him Frank, where we would never call Wittgenstein “Ludwig” (let alone Luki) or even Russell “Bertie”.

In a handful of years, he founded a couple of subfields of mathematics, a couple in economics, produced the greatest theory in epistemology, and one of the most radical notions in ethics, which has upended my life. A Bayesian before Jeffreys and de Finetti; a longtermist before the Bomb.

But even massive nerds haven’t heard of him. I just polled my philosophy degree mates, and only 2 of 10 had heard of him, and only because of his link to the Cult Wittgenstein.

This is probably because he’s too mathematical for the novelist-journalist-historian-biographer complex that determines popular stature. 3 Someday I will get around to writing an Explain Like I’m Five for all 16 of his publications.

Enough ass-kissing, what did he do?

## Achievements

- Ramsey’s theorem (founded Ramsey Theory)
  - groundwork in **extremal graph theory**.
- Ramsey Sentences
- Ramsey’s Maxim,
- Ramsey’s Theory of Truth,
- Ramsey Pricing
- Ramsey’s Theory of Preference
- Ramsey Theory of optimal saving
- Ramsey Theory of optimal taxation
  - Moral discount rate is zero.

- Ramsey's Theory of Probability
- Ramsey Numbers
- Ramsey Cardinals
- Ramsification
- Translator of last resort of Wittgenstein's *Tractatus* (aged 18)

2

### **Humanism, pragmatism, altruism, fun**

But alongside these is his philosophy-of-life or anyway his demeanour. Keynes again:

His bulky Johnsonian frame, his spontaneous gurgling laugh, the simplicity of feelings and reactions, half-alarming sometimes and occasionally almost cruel in their directness and literalness, his honesty of mind and heart, his modesty, and the amazing, easy efficiency of the intellectual machine which ground away behind his wide temples and broad, smiling face, have been taken from us at the height of their excellence and before their harvest of work and life could be gathered in.

He models a genius without pretensions, without cynicism, without abuse or neglect, and without self-conscious tragedy. This is an important model, if only to show up the many people who think that suffering is deeper than fun and intelligence is a great burden. You're no Frank Ramsey.

"My picture of the world is drawn in perspective, and not like a model to scale. The foreground is occupied by human beings and the stars are all as small as threepenny bits... I apply my perspective not merely to space but also to time. In time the world will cool and everything will die; but that is a long time off still, and its present value at compound discount is almost nothing. Nor is the present less valuable because the future will be blank. Humanity, which fills the foreground of my picture, I find interesting and on the whole admirable. I find, just now at least, the world a pleasant and exciting place. You may find it depressing; I am sorry for you, and you despise me. But the world is not in itself good or bad; it is just that it thrills me but depresses you. On the other hand, I pity you with reason, because it is pleasanter to be thrilled than to be depressed, and not merely pleasanter but better for all one's activities."

Although the study of classics and mathematics is very valuable, more valuable, perhaps, than other games like chess, yet it is difficult to see that proficiency at it is a sufficient reason why a man should not do his share of the world's work, give something to his fellows in exchange for the meat and drink they give him... Geometry is doubtless [a] more divine amusement... but what excuse is that...

[to] not give up his amusement to save his creatures from their present miserable condition?"

## Influence

Really not much directly. Dying young wasn't good for the spread of his results (unlike the musician case) and he quickly became obscure. So we had to rediscover half of his work. But this means we can make an excellent tragic estimate of how far ahead of the curve he was: 11 years ahead of de Finetti; 28 years ahead of Boiteux; 44 years ahead of Diamond and Mirrlees... 4

- Aged 26, he was Wittgenstein's (aged 40) PhD advisor. He inspired some amount of Wittgenstein's *Philosophical Investigations*; opinions vary about whether this was as Hume to his Kant or Bernoulli to de l'Hôpital.

## Misak

His instincts, in all parts of his life, were straightforward and directed to the facts.

Roy Harrod: "Ramsey's intellectual process is at white heat; but the style is delightfully cool, like that of some old naturalist taking one for a ramble in the country and making desultory observations."

Sporty despite being ungainly and hyperintellectual. Heroic

## Winchester

- The seventy scholarship boys lived in a damp and cold fourteenth-century building. The ch...
- The war had a deleterious effect on the food, which at Winchester was at the best of times...
- [Interesting that the "scholars" [scholarship boys] were higher status than the "commoners"]

The primary issue was the length of the day, which officially started at 6:45 in the morning and went straight through to 8:45 in the evening. Lessons began at 7 am, and late risers did without their cup of tea and two weevily biscuits. But the junior boys had a much more arduous time of it than the official story let on. As one of the youngest, it was Frank's duty to get up at the first sound of the bell at 6 am and 'call' or wake up the dormitory. - Frank took life seriously and did not like being made fun of - The police themselves went out on strikes during 1918–19. A revolutionary socialism now seemed a real possibility. - Charlie would no doubt turn in his grave if he knew that the only reason anyone still reads his history of the Wilson family is to find out about those two left-wing nephews. - "The reason people say this is the happiest time of one's life is simply that... most people for e.g. Foot will never again possess despotic power as he does now. He can be witness judge and executioner all in one; he can abuse juniors as he will never-again be able to

abuse people . . . He can go about imagining he is upholding the foundations of College and talk rot about prefectorial dignity and people being above themselves and can beat people as he did Asquith for being ‘solitary’”

## Cambridge

“We really live in a great time for thinking”

He would become the singular mind who could engage each of those great thinkers—Keynes, Russell, Moore, Wittgenstein—on their own terms. There would be simply no one else who could do that, including these four themselves. At the least, Keynes and Moore weren’t up to Russell’s logical skills and Wittgenstein, Russell, and Moore weren’t up to Keynes’s skill in economics.

## Of his time

Although he established much that will outlast you and me and our memory, he was also caught up in the manias of his day. He was a staunch socialist, caught the anti-masturbation fad, and devoted months and months of his short life to being psychoanalysed.

(He was also in a slightly mishandled open relationship with his wife, which seems more prescient than Bloomsbury-quaint.)

I suppose I am equally timebound in some way or other. (Music taste, most likely.) I suppose there’s no shame in it.

His economics offer a neat rebuttal to naive accusations about the field’s polarised ideology: his theories of tax and saving are still central to the field, despite being about modelling active governments and a wealth tax. (Never mind Samuelson and Arrow, never mind the vast Keynesian contingent.) More generally he is a happy case of a strong ideology tamed by mathematics, of a productive partisan.

<h3>Was Ramsey the first longtermist?</h3>

<div>

Or the first quant longtermist: zero discount, to treat the future as you would want to [Sidgwick] ([https://www.libraryathena.com/the-methods-of-ethics\\_69](https://www.libraryathena.com/the-methods-of-ethics_69)) already got the point

<blockquote>the time at which a man exists cannot affect the value of his happiness from

<!-- -->

<a href="{{ben}}>Bentham was positive discount</a> (fn6). Mill thought about personal t

</div>

<!-- -->

<h3>Karl Sabbagh on Ramsey</h3>

<div>

Someone wrote a tiny biography of Ramsey before Misak. He is fairly obsessed with Frank

<!-- -->

His teen diary:

```
<blockquote>Wonder what I shall do for schedule B. Feel inclined to geometry but I'm suc  
Went to Mikado. It is a glorious show<br><br>  
Did accounts. Have spent £86 5s since beginning of academical year, reckon ought to do p  
Really angry with myself re sex. Woe unto them that desire things that give no satisfact  
I do hope I'm not impugning my health but haven't the guts to talk to a doctor.  
</blockquote>  
</div>  
{% include ramsey/foots.md %}
```

Peirce

Gavin

2022-02-01

# von Neumann

Gavin

2022-01-01

An awaited book; in fact I awaited it before I knew it was being written. Here is one of the most important people to ever live, and what notice do we take? Before now: One bad old biography (and one-third of another) and many gigantic maths monographs. Such yawning gaps come from historians and biographers being obsessed with artists instead - consider the nine Jane Austen biographies published in the last 11 years - and our scientists being inarticulate at best. unable or unwilling to stand up for themselves, and unrepresented by the chattering classes. 1

It is incredibly difficult to cover everything von Neumann did - everything he did for the first time in history - even just everything with vast practical consequences which are still felt 60 years later.

- Chapter 2: fixing set theory where Hilbert and Russell failed
- Chapter 3: unifying matrix and wave theory where Dirac bodge and others failed
- Chapter 4: solution to a profound engineering challenge which changed the world forever
- ...

Great philosophers get several kinds of books written about them - two are the Life (which gossips about their upbringing and vices), and the intellectual biography (which actually tries to explain and show the development of their ideas). Bhattacharya's is more like the latter plus a smattering of parties, fast cars, and intellectual bitching.

Hodges is, in 600 pages, just able to enumerate Turing's achievements. Bhattacharya, in 284, is not even vaguely able to do this for vN. e.g. Almost no mention of his great work in group theory.

---

Very incomplete list of von Neumann's achievements:

- Foundations of maths: Paradox-free foundation of set theory with classes (superceded Russell)

- Physics: Unification of matrix mechanics and wave mechanics (superceded Dirac)
- Physics:: proof of the Ergodic hypothesis
- Lots of group theory, chiefly operator algebras
- Foundations of QM: axiomatisation of QM, unified wave and matrix mechanics.
- Physics: Clarified the measurement problem (for the first time?)
- Physics:: Central work on the Copenhagen interpretation
- Physics / logic Founded quantum logic
- Economics: Proved existence and uniqueness of general equilibrium
- Physics: Much-misunderstood constraint on all hidden variable theories. Maybe gappy.
- linear programming: duality and the first interior point method.
- Fluid dynamics: Fat Man implosion lens design. Discovery of the airburst efficiency. Many solutions in blast waves.
- Hardware engineering: Redesigned the ENIAC to be the first stored program computer
- Computer engineering: Earliest partial design of a modern computer. Lifted lots from Mauchly and Eckert (uncredited) but greatly superceded them.
- Patent busting on the digital computer design. Free for all.
- Minimax and dozens of central results in game theory
- Founded utility theory
- Marrying neuroscience and computer science forever
- Founded automata theory
- Intelligence explosion as x-risk
- ...

Bhattacharya covers about half of these.

---

- The most important question in all of education: How did Hungary produce so many geniuses? Why did they stop? The second has an obvious answer (the Holocaust), but the first is tricky. Theories of Jewish excellence do not suffice: why Hungary instead of Poland (ten times larger population), Czechoslovakia, Britain? von Neumann's own answer was the empire's weird mix of 1) tolerance and rewards for Jewish people, while 2) still being

extremely volatile and so making them uncertain how long this would last and so rushing to succeed.

- Bhattacharya's informalisation of the technical results here is impressive. At least one fuckup though: on p112 he confuses completeness for correctness.
  - At one point AB ties the Hilbert and Gödel work to modernism. Modernist mathematics, the rejection of the past, the flight into abstraction and rigour. As if this was a general spirit. I don't know how to evaluate this idea.
  - Sad to hear that a heavily modified ENIAC executed a stored program two months before the Manchester Baby. I hate to see the Man win over the garage nerds.
  - Nash is nasty, well before he goes psychotic (self-aggrandising, straw Vulcan, racist). He makes von Neumann look soft and warm.
  - Lovelace is not the first programmer. Klári von Neumann has a much better claim (if we insist on ignoring Babbage).
  - So many brilliant people here, and far more obscure than JvN. Shapley, Barricelli, Collbohm, Goldstine, Harsanyi, McCarthy, Adele and Klári...
- 

## Err

We tend to deify people, and they never deserve it. What did von Neumann get wrong?

### Mutually Assured Destruction

It's not obvious that this was a mistake - we're still here, MAD is a strong reason not to intentionally nuke people. But the sheer number of near misses and the overall estimate of 0.1% annual state risk, should make us think that the strategy was actually poor, that we are walking selection bias. The less obvious response is that he knew all that and was trading some existential risk to block the Soviet Empire's anti-human practices from taking over. Since this argument also works for the Soviets, or for any value system which values itself, he seems to have settled for an appalling equilibrium. Tragedy of the value lock-in commons.

VN wanted cooperation, wanted a long life for humanity. But he couldn't trust enough not to escalate. The true altruist cannot afford to cooperate simply.

### First strike on the Soviets

If you say 'why not bomb them tomorrow', I say, why not today? If you say 'today at five o'clock', I say why not one o'clock?

(He recanted this a couple of years later.)

The mistake was twofold: to assume that the Soviets would continue growing, and to assume that the nuclear taboo would not hold. That taboo, that tradition is one of the most precious things in the world, and almost nothing is worth breaking it. To which you reply: 100 million people are not worth it? To which I can only apologise and suggest that 100 million are not worth 300 million.

### **Trusting Klaus Fuchs**

He actually handed the Soviets a new nuke design through the infamous Fuchs.

### **Nonerror: “Proof” of no hidden variables**

The conventional view is that von Neumann screwed up his no-hidden-variables proof, claimed to have shown the impossibility of hidden variables, and that this convinced everyone until Bell came along and exposed the error (30 years after Grete Hermann did it and was ignored). But this misrepresents the proof, which just says that a hidden variables theory will have to have a certain weird structure (which Bohmian mechanics does).

### **Targeting Kyoto**

I don't know if a nuclear strike on Japan was ultimately for the best (considering the appalling toll of the Pacific theatre on both sides, the likely larger toll of taking Honshū, and the second-order effects of showing the world that everything had changed). But that they were civilian strikes seems completely gratuitous. Striking Kyoto, the spiritual centre, in particular seems incredibly high risk.

### **Nonerror: The brain is digital**

People act like he was naive about the brain as computer, but he just wasn't: the brain can *prima facie* be considered as a digital computer. However, upon further reflection, some elements of analog computing (e.g., the chemistry) will also become relevant in understanding the functioning of the brain.

### **The von Neumann bottleneck**

The world standard architecture for computers leads to a huge waste of CPU cycles, waiting for memory. This wasn't such a big deal in the 50s, but CPU performance has masssively outpaced bus bandwidth over the last 70 years.

### **Against high-level programming**

'von Neumann opposed the development of assemblers and high-level language compilers. He preferred to employ legions of human programmers (mostly low-paid graduate students) to hand-assemble code into machine language. "It is a waste of a valuable scientific computing instrument", von Neumann reportedly said, "to use it to do clerical work."'

## **Various dumb personal risks**

He did not live like someone who understood expected utility and hyperbolic discounting. He ate way too much, drove incredibly badly, was an easy mark for salesmen, pissed off his wife by leching. He spent a lot of time travelling to government meetings. He let others profit from his inventions. These imply irrationality - or a surprising lack of interest in his own wealth, longevity, time use, or marriage. This post collects other apparently bad decisions.

{% include jvn/foots.html %}

# History in the making

Gavin

2012-03-09

{% assign d = “[https://en.wikipedia.org/wiki/Kurt\\_Diebner](https://en.wikipedia.org/wiki/Kurt_Diebner)” %} {% assign t = “[http://sillok.history.go.kr/id/kca\\_10402008\\_004](http://sillok.history.go.kr/id/kca_10402008_004)” %}

“historians in history” “historical documents about historians” Herodotus primary sources

In 1404, King Taejong fell from his horse during a hunting expedition. Embarrassed, looking to his left and right, he commanded, “Do not let the historian find out about this.” To his disappointment, the historian accompanying the hunting party included these words in the annals, in addition to a description of the king’s fall.

---

Diebner: I wonder whether there are microphones installed here?  
Heisenberg: Microphones installed? (laughing) Oh no, they’re not as cute as all that. I don’t think they know the real Gestapo methods; they’re a bit old fashioned in that respect.

---

Bedding ceremony

Consummations witnessed in Sweden, Scotland, Iceland <https://brill.com/view/title/13763>

# The greatest returns on investment ever

Gavin

2022-02-03

```
<h3>Boundary work</h3>
<div>
    accruing to the original entity
    not the Imperial Family of Japan
    not inheritance??
    but yes to corps???
```

no chaining investments Has to be realised gains Has to be financial investment (not labour, not ideas) in 2022 dollars lean towards major outcomes (>\$10m) ex post, obviously. This isn't to call them geniuses.

<!--

```
-->
<h3>Ceiling</h3>
<div>
</div>
```

**1 million x. Ceiling**

**260,000x. Ground-floor Bitcoin mining**

2009 address 40 BTC (\$391,055)

Mining this would have cost about \$1.20 in energy in 2010. Add in \$0.30 for the amortized PC. (This might not be realised, just shuffling around, but good enough.)

**??,000x. Dutch East India Company**

1602. Dutch govt. \$7.8T at peak realised 1799 Mixture of trade and plunder monopoly over the Asian trade. For a time in the seventeenth century, it was able to monopolise the trade in nutmeg, mace, and cloves Taiwan, Mauritius, South Africa, Jakarta, Ceylon, Bengal, Hanoi

### **25,000x. Wardian cases**

UK Government (Kew Gardens): Wardian cases for tea and quinine and rubber  
20,000 tea plants smuggled out of Shanghai to begin the tea plantations of Assam  
in 1849. world's largest tea-growing region £10bn a year \* 65 years = 150bn  
1860, quinine from Brazil basis for empire £6.4m a year \* 100 years = £1bn  
1870, rubber from Brazil Ceylon and Malaya later the main source for British  
WW2 military £100bn Retired in 1962.

### **21,000x. Bletchley.**

1938. UK Government: £3000 (£137,962) for five years of Alan Turing's best work  
Total military budget was £3bn a year. Call it 0.5% = £15m

£6,000 (£392,000) for the grounds Returns War shortened by 2 years £2.6  
trillion = 14 million lives at SVOL £1.83m / 10 About 2000 ships saved,  
10 million tons About a third of all ships \$1 trillion April 1945 July 1941  
= 3.7 years of cracks

### **5000x. Apple.**

1977. Mike Markkula: \$80,000 (\$368,055) for one-third of Apple. Now worth  
\$1.5bn But had he held all the way and sold now, 7.8 million x

### **1000x floor: stock average**

Dow Jones Industrial Average May 26, 1896 = **40.94** 36,000

### **TODO**

Berkshire Hathaway

Opium?

Escobar coke

### **????x. Conquest of the new world**

1530. Charles V Pizarro, Cortes

I received little help from historians. Almost none of them think in these terms;  
few pay much heed to two orders of magnitude, or any.

# Legacy and the memory of legacy

Gavin

2022-02-20

```
{% assign damn="https://en.wikipedia.org/wiki/Damnatio_memoriae" %}  
{% assign wtb = "https://www.goodreads.com/book/show/33517544-the-will-to-battle" %} {% assign flor = "https://www.jstor.org/stable/43446248" %}  
{% assign hero = "https://en.wikipedia.org/wiki/Herostratus" %} {% assign murder = "https://www.theatlantic.com/national/archive/2014/05/mass-murder-should-not-be-a-ticket-to-fame/371599/" %} {% assign ahk = "https://smarthistory.org/house-altar-depicting-akhenaten-nefertiti-and-three-of-their-daughters/" %} {% assign co = "https://en.wikipedia.org/wiki/Commodus#/media/File:R%C3%B6merr-09-30_008.jpg" %} {% assign singer = "https://www.thelifeyoucansave.org/" %}
```

A grave punishment in ancient Rome was *damnatio memoriae*: being written out of history. Ada Palmer's extremely melodramatic and for all I know accurate portrayal:

[the damned person is] neither slim nor mighty, stooped nor noble, just a shape... Somewhere in a dusty archive a baptismal registry records some Hildebrand, and, when that dry page molder... I can't look, I can't! Behind the shades, the broad gray plain, that sea of shapeless gloom extending on and on... all forgotten souls, minds empty of memory, smeared one into another... to this absolute dissolution Caesar damns his enemies... Not me! I will never let you take me! I will carve my memory into history, by work, by force, by guile, in swathes of blood and ashes if I must!

Supposedly this remained an effective policy in the Renaissance:

In 1343... the Florentine republic that replaced the Duke of Athens ordered all memory of him and his rule erased, and all images and mementos of him destroyed immediately... In addition to the official punishments, a crowd of citizens stormed the government palace in order to burn archival documents... A crowd also cannibalized two of the duke's supporters in a particularly brutal form of bodily *damnatio memoriae* that seems to have emulated the corpse abuse practised on hated emperors in ancient Rome.

Sometimes people even seem to prefer being lied about and demonised to being forgotten.

I cannot understand this at all, and (oddly for me) I don't want to. It just doesn't seem like a big deal. The pain and abuse of power preceding your expurgation is overwhelmingly more important.

I know why you'd do it to ideological opponents - to hide your crimes, to manage competing ideologies and pre-empt martyrs. So I understand the negationism of Seti and Stalin and the rest. (Actually, how often did it work? Lots of damned people are now more famous than their damners, an ur-Streisand effect. But *maybe* some cases were done so well that I will never know the numerator here.)

Then there's a sensible kind, which just attempts to remove the incentive for people to commit infamous crimes just for the sake of fame. (Our media merrily incentivise murder all the time.)

So I'm instead mocking the reaction of the target to posthumous punishment. Fearing *damnatio memoriae* is an ultimate kind of wounded pride. Men who appear to value being remembered more than life or anything. This seems related to the naive idea of 'living on' through your descendants.

I want to shake them. "Look man, I know we're all status-obsessed, but some of us try to earn status by doing things. Look man, I know death sucks ass, but using history as a consolation prize is pathetic."

---

My stepfather 1 died a few years ago. He was a nice man, but comically taciturn. I was a nice lad, but comically shy. We probably had about 5 serious conversations in 5 years.

One of them concerned his final rest. He told me that he wanted absolutely no monument, no gravestone and no plaque. He told me that it was meaningless and greedy to cling to things when you have no fingers. That he'd had his share of the world. God wasn't in it. When his sons in turn were gone, he wanted to disturb the waters no more. We dumped his ashes - the ashes - at sea.

Let's say he wasn't exceptional in this, that the mania for legacy has declined between Imperium and now. (Rather than being sublimated somehow.) A huge change. Meaning, relocated from public stature to private experience. Status, bounded by one life and one small group of people. Honour, a matter of living peacefully, tidying up after yourself, and turning off the light.

---

There is a version of this which would scare me - if anyone ever hated me enough to do it. Call it *damnatio opera*.

This is not the pathetic, primitive damnation of having your name chiselled off plinths and deleted from databases. "Ow my status!!" This is the undoing of everything good you have *done*. Your children eliminated, certainly. But also an

unkindness to every person you've been kind to. An opposite murder for every life you save. Your parents' pride undone. All your writing, bit-rotted. All your arguments, refuted. All your charity seized. All that you inspired pruned. All happy memories spoiled or repressed.

Maybe people still know that you existed. But so what?

Something less than a stepfather but more than my mother's boyfriend idk.

Most of mathematics is forever wild

Remember! Most strings are incompressible, most reals uncomputable, most theorems unprovable, most programs undecidable.

- Gwern

Are most antiderivatives nonelementary??

Most angles of  $\sin x$  have no closed-form solution (because of the Abel–Ruffini theorem) nah. true for rational form, not exponential i form