Best of LessWrong: January 2015

1. CFAR fundraiser far from filled; 4 days remaining
2. The Importance of Sidekicks
3. 2014 Survey Results
4. Bill Gates: problem of strong AI with conflicting goals "very worthy of study and time"
5. Behavior: The Control of Perception
6. Elon Musk donates $10M to the Future of Life Institute to keep AI beneficial
7. Immortality: A Practical Guide
8. An Introduction to Control Theory
9. 'Dumb' AI observes and manipulates controllers
10. Overpaying for happiness?

Best of LessWrong: January 2015

1. CFAR fundraiser far from filled; 4 days remaining
2. The Importance of Sidekicks
3. 2014 Survey Results
4. Bill Gates: problem of strong AI with conflicting goals "very worthy of study and time"
5. Behavior: The Control of Perception
6. Elon Musk donates $10M to the Future of Life Institute to keep AI beneficial
7. Immortality: A Practical Guide
8. An Introduction to Control Theory
9. 'Dumb' AI observes and manipulates controllers
10. Overpaying for happiness?

CFAR fundraiser far from filled; 4 days remaining

We're 4 days from the end of our matching fundraiser, and still only about 1/3rd of the way to our target (and to the point where pledged funds would cease being matched).

If you'd like to support the growth of rationality in the world, do please consider donating, or asking me about any questions/etc. you may have. I'd love to talk. I suspect funds donated to CFAR between now and Jan 31 are quite high-impact.

As a random bonus, I promise that if we meet the $120k matching challenge, I'll post at least two posts with some never-before-shared (on here) rationality techniques that we've been playing with around CFAR.

The Importance of Sidekicks

[Reposted from my personal blog.]

Mindspace is wide and deep. “People are different” is a truism, but even knowing this, it’s still easy to underestimate.

I spent much of my initial engagement with the rationality community feeling weird and different. I appreciated the principle and project of rationality as things that were deeply important to me; I was pretty pro-self improvement, and kept tsuyoku naritai as my motto for several years. But the rationality community, the people who shared this interest of mine, often seemed baffled by my values and desires. I wasn’t ambitious, and had a hard time wanting to be. I had a hard time wanting to be anything other than a nurse.

It wasn’t until this August that I convinced myself that this wasn’t a failure in my rationality, but rather a difference in my basic drives. It’s around then, in the aftermath of the 2014 CFAR alumni reunion, that I wrote the following post.

> I don’t believe in life-changing insights (that happen to me), but I think I’ve had one–it’s been two weeks and I’m still thinking about it, thus it seems fairly safe to say I did.
>
> At a CFAR Monday test session, Anna was talking about the idea of having an “aura of destiny”–it’s hard to fully convey what she meant and I’m not sure I get it fully, but something like seeing yourself as you’ll be in 25 years once you’ve saved the world and accomplished a ton of awesome things. She added that your aura of destiny had to be in line with your sense of personal aesthetic, to feel “you.”
>
> I mentioned to Kenzi that I felt stuck on this because I was pretty sure that the combination of ambition and being the locus of control that “aura of destiny” conveyed to me was against my sense of personal aesthetic.
>
> Kenzi said, approximately [I don't remember her exact words]: “What if your aura of destiny didn’t have to be those things? What if you could be like…Samwise, from Lord of the Rings? You’re competent, but most importantly, you’re *loyal* to Frodo. You’re the reason that the hero succeeds.”
>
> I guess this isn’t true for most people–Kenzi said she didn’t want to keep thinking of other characters who were like this because she would get so insulted if someone kept comparing her to people’s sidekicks–but it feels like now I know what I am.
>
> So. I’m Samwise. If you earn my loyalty, by convincing me that what you’re working on is valuable and that you’re the person who should be doing it, I’ll stick by you whatever it takes, and I’ll *make sure* you succeed. I don’t have a Frodo right now. But I’m looking for one.

It then turned out that quite a lot of other people recognized this, so I shifted from “this is a weird thing about me” to “this is one basic personality type, out of many.” Notably, Brienne wrote the following comment:

> Sidekick” doesn’t *quite* fit my aesthetic, but it’s extremely close, and I feel it in certain moods. Most of the time, I think of myself more as what TV tropes would call a “dragon”. Like the Witch-king of Angmar, if we’re sticking of LOTR. Or Bellatrix Black. Or Darth Vader. (It’s not my fault people aren’t willing to give the good guys dragons in literature.)
>
> For me, finding someone who shared my values, who was smart and rational enough for me to trust him, and who was in a much better position to actually accomplish what I most cared about than I imagined myself ever being, was the best thing that could have happened to me.

She also gave me what’s maybe one of the best and most moving compliments I’ve ever received.

> In Australia, something about the way you interacted with people suggested to me that you help people in a completely free way, joyfully, because it fulfills you to serve those you care about, and not because you want something from them… I was able to relax around you, and ask for your support when I needed it while I worked on my classes. It was really lovely… The other surprising thing was that you seemed to act that way with everyone. You weren’t “on” all the time, but when you were, everybody around you got the benefit. I’d never recognized in anyone I’d met a more diffuse service impulse, like the whole human race might be your master. So I suddenly felt like I understood nurses and other people in similar service roles for the first time.

Sarah Constantin, who according to a mutual friend is one of the most loyal people who exists, chimed in with some nuance to the Frodo/Samwise dynamic: “Sam isn’t blindly loyal to Frodo. He makes sure the mission succeeds even when Frodo is fucking it up. He stands up to Frodo. And that’s important too.”

Kate Donovan, who also seems to share this basic psychological makeup, added “I have a strong preference for making the lives of the lead heroes better, and very little interest in ever being one.”

Meanwhile, there were doubts from others who didn’t feel this way. The “we need heroes, the world needs heroes” narrative is especially strong in the rationalist community. And typical mind fallacy abounds. It seems easy to assume that if someone wants to be a support character, it’s because they’re insecure–that really, if they believed in themselves, they would aim for protagonist.

I don’t think this is true. As Kenzi pointed out: “The other thing I felt like was important about Samwise is that his self-efficacy around his particular mission wasn’t a detriment to his aura of destiny – he did have insecurities around his ability to do this thing – to stand by Frodo – but even if he’d somehow not had them, he still would have been Samwise – like that kind of self-efficacy would have made his essence *more* distilled, not less.”

Brienne added: “Becoming the hero would be a personal tragedy, even though it would be a triumph for the world if it happened because I surpassed him, or discovered he was fundamentally wrong.”

Why write this post?

Usually, “this is a true and interesting thing about humans” is enough of a reason for me to write something. But I’ve got a lot of other reasons, this time.

I suspect that the rationality community, with its “hero” focus, drives away many people who are like me in this sense. I’ve thought about walking away from it, for basically that reason. I could stay in Ottawa and be a nurse for forty years; it would fulfil all my most basic emotional needs, and no one would try to change me. Because oh boy, have people tried to do that. It’s really hard to be someone who just wants to please others, and to be told, basically, that you’re not good enough–and that you owe it to the world to turn yourself ambitious, strategic, Slytherin.

Firstly, this is mean regardless. Secondly, it’s not true.

Samwise was important. So was Frodo, of course. But Frodo needed Samwise. Heroes need sidekicks. They can function without them, but function a lot better with them. Maybe it’s true that there aren’t enough heroes trying to save the world. But there sure as hell aren’t enough sidekicks trying to help them. And there especially aren’t enough talented, competent, awesome sidekicks.

If you’re reading this post, and it resonates with you… Especially if you’re someone who has felt unappreciated and alienated for being different… I have something to tell you. You count. You. Fucking. Count. You’re needed, even if the heroes don’t realize it yet. (Seriously, heroes, you should be more strategic about looking for awesome sidekicks. AFAIK only Nick Bostrom is doing it.) This community could use more of you. Pretty much *every* community could use more of you.

I’d like, someday, to live in a culture that doesn’t shame this way of being. As Brienne points out, “Society likes *selfless* people, who help everybody equally, sure. It’s socially acceptable to be a nurse, for example. Complete loyalty and devotion to “the hero”, though, makes people think of brainwashing, and I’m not sure what else exactly but bad things.” (And not all subsets of society even accept nursing as a Valid Life Choice.) I’d like to live in a world where an aspiring Samwise can find role models; where he sees awesome, successful people and can say, “yes, I want to grow up to be that.”

Maybe I can’t have that world right away. But at least I know what I’m reaching for. I have a name for it. And I have a Frodo–Ruby and I are going to be working together from here on out. I have a reason not to walk away.

2014 Survey Results

Thanks to everyone who took the 2014 Less Wrong Census/Survey. Extra thanks to Ozy, who did a lot of the number crunching work.

This year's results are below. Some of them may make more sense in the context of the original survey questions, which can be seen here. Please do not try to take the survey as it is over and your results will not be counted.

I. Population

There were 1503 respondents over 27 days. The last survey got 1636 people over 40 days. The last four full days of the survey saw nineteen, six, and four responses, for an average of about ten. If we assume the next thirteen days had also gotten an average of ten responses - which is generous, since responses tend to trail off with time - then we would have gotten about as many people as the last survey. There is no good evidence here of a decline in population, although it is perhaps compatible with a very small decline.

II. Demographics

Sex
Female: 179, 11.9%
Male: 1311, 87.2%

Gender
F (cisgender): 150, 10.0%
F (transgender MtF): 24, 1.6%
M (cisgender): 1245, 82.8%
M (transgender FtM): 5, 0.3%
Other: 64, 4.3%

Sexual Orientation
Asexual: 59, 3.9%
Bisexual: 216, 14.4%
Heterosexual: 1133, 75.4%
Homosexual: 47, 3.1%
Other: 35, 2.3%

[This question was poorly worded and should have acknowledged that people can both be asexual and have a specific orientation; as a result it probably vastly undercounted our asexual readers]

Relationship Style
Prefer monogamous: 778, 51.8%
Prefer polyamorous: 227, 15.1%
Uncertain/no preference: 464, 30.9%
Other: 23, 1.5%

Number of Partners
0: 738, 49.1%
1: 674, 44.8%
2: 51, 3.4%
3: 17, 1.1%
4: 7, 0.5%
5: 1, 0.1%
Lots and lots: 3, 0.2%

Relationship Goals
Currently not looking for new partners: 648, 43.1%
Open to new partners: 467, 31.1%
Seeking more partners: 370, 24.6%

[22.2% of people who don't have a partner aren't looking for one.]

Relationship Status
Married: 274, 18.2%
Relationship: 424, 28.2%
Single: 788, 52.4%

[6.9% of single people have at least one partner; 1.8% have more than one.]

Living With
Alone: 345, 23.0%
With parents and/or guardians: 303, 20.2%
With partner and/or children: 411, 27.3%
With roommates: 428, 28.5%

Children
0: 1317, 81.6%
1: 66, 4.4%
2: 78, 5.2%
3: 17, 1.1%
4: 6, 0.4%
5: 3, 0.2%
6: 1, 0.1%
Lots and lots: 1, 0.1%

Want More Children?
Yes: 549, 36.1%
Uncertain: 426, 28.3%
No: 516, 34.3%

[418 of the people who don't have children don't want any, suggesting that the LW community is 27.8% childfree.]

Country
United States, 822, 54.7%
United Kingdom, 116, 7.7%
Canada, 88, 5.9%
Australia: 83, 5.5%
Germany, 62, 4.1%
Russia, 26, 1.7%
Finland, 20, 1.3%
New Zealand, 20, 1.3%
India, 17, 1.1%
Brazil: 15, 1.0%
France, 15, 1.0%
Israel, 15, 1.0%

Lesswrongers Per Capita
Finland: 1/271,950
New Zealand: 1/223,550
Australia: 1/278,674
United States: 1/358,390
Canada: 1/399,545
Israel: 1/537,266
United Kingdom: 1/552,586
Germany: 1/1,290,323
France: 1/ 4,402,000
Russia: 1/ 5,519,231
Brazil: 1/ 13,360,000
India: 1/ 73,647,058

Race
Asian (East Asian): 59. 3.9%
Asian (Indian subcontinent): 33, 2.2%
Black: 12. 0.8%
Hispanic: 32, 2.1%
Middle Eastern: 9, 0.6%
Other: 50, 3.3%
White (non-Hispanic): 1294, 86.1%

Work Status
Academic (teaching): 86, 5.7%
For-profit work: 492, 32.7%
Government work: 59, 3.9%
Homemaker: 8, 0.5%
Independently wealthy: 9, 0.6%
Nonprofit work: 58, 3.9%
Self-employed: 122, 5.8%
Student: 553, 36.8%
Unemployed: 103, 6.9%

Profession
Art: 22, 1.5%
Biology: 29, 1.9%
Business: 35, 4.0%
Computers (AI): 42, 2.8%
Computers (other academic): 106, 7.1%
Computers (practical): 477, 31.7%
Engineering: 104, 6.1%
Finance/Economics: 71, 4.7%
Law: 38, 2.5%
Mathematics: 121, 8.1%

Medicine: 32, 2.1%
Neuroscience: 18, 1.2%
Philosophy: 36, 2.4%
Physics: 65, 4.3%
Psychology: 31, 2.1%
Other: 157, 10.2%
Other “hard science”: 25, 1.7%
Other “social science”: 34, 2.3%

Degree
None: 74, 4.9%
High school: 347, 23.1%
2 year degree: 64, 4.3%
Bachelors: 555, 36.9%
Masters: 278, 18.5%
JD/MD/other professional degree: 44, 2.9%
PhD: 105, 7.0%
Other: 24, 1.4%

III. Mental Illness

535 answer “no” to all the mental illness questions. Upper bound: 64.4% of the LW population is mentally ill.
393 answer “yes” to at least one mental illness question. Lower bound: 26.1% of the LW population is mentally ill. Gosh, we have a lot of self-diagnosers.

Depression
Yes, I was formally diagnosed: 273, 18.2%
Yes, I self-diagnosed: 383, 25.5%
No: 759, 50.5%

OCD
Yes, I was formally diagnosed: 30, 2.0%
Yes, I self-diagnosed: 76, 5.1%
No: 1306, 86.9%

Autism spectrum
Yes, I was formally diagnosed: 98, 6.5%
Yes, I self-diagnosed: 168, 11.2%
No: 1143, 76.0%

Bipolar
Yes, I was formally diagnosed: 33, 2.2%
Yes, I self-diagnosed: 49, 3.3%
No: 1327, 88.3%

Anxiety disorder
Yes, I was formally diagnosed: 139, 9.2%
Yes, I self-diagnosed: 237, 15.8%
No: 1033, 68.7%

BPD
Yes, I was formally diagnosed: 5, 0.3%
Yes, I self-diagnosed: 19, 1.3%
No: 1389, 92.4%

[Ozy says: RATIONALIST BPDERS COME BE MY FRIEND]

Schizophrenia
Yes, I was formally diagnosed: 7, 0.5%
Yes, I self-diagnosed: 7, 0.5%
No: 1397, 92.9%

IV. Politics, Religion, Ethics

Politics
Communist: 9, 0.6%
Conservative: 67, 4.5%
Liberal: 416, 27.7%
Libertarian: 379, 25.2%
Social Democratic: 585, 38.9%

[The big change this year was that we changed "Socialist" to "Social Democratic". Even though the description stayed the same, about eight points worth of Liberals switched to Social Democrats, apparently more willing to accept that label than "Socialist". The overall supergroups Libertarian vs. (Liberal, Social Democratic) vs. Conservative remain mostly unchanged.]

Politics (longform)
Anarchist: 40, 2.7%
Communist: 9, 0.6%
Conservative: 23, 1.9%
Futarchist: 41, 2.7%
Left-Libertarian: 192, 12.8%
Libertarian: 164, 10.9%
Moderate: 56, 3.7%
Neoreactionary: 29, 1.9%
Social Democrat: 162, 10.8%
Socialist: 89, 5.9%

[Amusing politics answers include anti-incumbentist, having-well-founded-opinions-is-hard-but-I've-come-to-recognize-the-pragmatism-of-socialism-I-don't-know-ask-me-again-next-year, pirate, progressive social democratic environmental liberal isolationist freedom-fries loving pinko commie piece of shit, republic-ist aka read the federalist papers, romantic reconstructionist, social liberal fiscal agnostic, technoutopian anarchosocialist (with moderate snark), whatever it is that Scott is, and WHY ISN'T THERE AN OPTION FOR NONE SO I CAN SIGNAL MY OBVIOUS OBJECTIVITY WITH MINIMAL EFFORT. Ozy would like to point out to the authors of manifestos that no one will actually read their manifestos except zir, and they might want to consider posting them to their own blogs.]

American Parties
Democratic Party: 221, 14.7%
Republican Party: 55, 3.7%
Libertarian Party: 26, 1.7%
Other party: 16, 1.1%
No party: 415, 27.6%
Non-Americans who really like clicking buttons: 415, 27.6%

Voting
Yes: 881, 58.6%
No: 444, 29.5%
My country doesn't hold elections: 5, 0.3%

Religion
Atheist and not spiritual: 1054, 70.1%
Atheist and spiritual: 150, 10.0%
Agnostic: 156, 10.4%
Lukewarm theist: 44, 2.9%
Deist/pantheist/etc.: 22,, 1.5%
Committed theist: 60, 4.0%

Religious Denomination
Christian (Protestant): 53, 3.5%
Mixed/Other: 32, 2.1%
Jewish: 31, 2.0%
Buddhist: 30, 2.0%
Christian (Catholic): 24, 1.6%
Unitarian Universalist or similar: 23, 1.5%

[Amusing denominations include anti-Molochist, CelestAI, cosmic engineers, Laziness, Thelema, Resimulation Theology, and Pythagorean. The Cultus Deorum Romanorum practitioner still needs to contact Ozy so they can be friends.]

Family Religion
Atheist and not spiritual: 213, 14.2%
Atheist and spiritual: 74, 4.9%
Agnostic: 154. 10.2%
Lukewarm theist: 541, 36.0%
Deist/Pantheist/etc.: 28, 1.9%
Committed theist: 388, 25.8%

Religious Background
Christian (Protestant): 580, 38.6%
Christian (Catholic): 378, 25.1%
Jewish: 141, 9.4%
Christian (other non-protestant): 88, 5.9%
Mixed/Other: 68, 4.5%
Unitarian Universalism or similar: 29, 1.9%
Christian (Mormon): 28, 1.9%
Hindu: 23, 1.5%'

Moral Views
Accept/lean towards consequentialism: 901, 60.0%
Accept/lean towards deontology: 50, 3.3%
Accept/lean towards natural law: 48, 3.2%
Accept/lean towards virtue ethics: 150, 10.0%
Accept/lean towards contractualism: 79, 5.3%

Other/no answer: 239, 15.9%

Meta-ethics
Constructivism: 474, 31.5%
Error theory: 60, 4.0%
Non-cognitivism: 129, 8.6%
Subjectivism: 324, 21.6%
Substantive realism: 209, 13.9%

V. Community Participation

Less Wrong Use
Lurker: 528, 35.1%
I've registered an account: 221, 14.7%
I've posted a comment: 419, 27.9%
I've posted in Discussion: 207, 13.8%
I've posted in Main: 102, 6.8%

Sequences
Never knew they existed until this moment: 106, 7.1%
Knew they existed, but never looked at them: 42, 2.8%
Some, but less than 25%: 270, 18.0%
About 25%: 181, 12.0%
About 50%: 209, 13.9%
About 75%: 242, 16.1%
All or almost all: 427, 28.4%

Meetups
Yes, regularly: 154, 10.2%
Yes, once or a few times: 325, 21.6%
No: 989, 65.8%

Community
Yes, all the time: 112, 7.5%
Yes, sometimes: 191, 12.7%
No: 1163, 77.4%

Romance
Yes: 82, 5.5%
I didn't meet them through the community but they're part of the community now: 79, 5.3%
No: 1310, 87.2%

CFAR Events
Yes, in 2014: 45, 3.0%
Yes, in 2013: 60, 4.0%
Both: 42, 2.8%
No: 1321, 87.9%

CFAR Workshop
Yes: 109, 7.3%
No: 1311, 87.2%

[A couple percent more people answered 'yes' to each of meetups, physical interactions, CFAR attendance, and romance this time around, suggesting the community is very very gradually becoming more IRL. In particular, the number of people meeting romantic partners through the community increased by almost 50% over last year.]

HPMOR
Yes: 897, 59.7%
Started but not finished: 224, 14.9%
No: 254, 16.9%

Referrals
Referred by a link: 464, 30.9%
HPMOR: 385, 25.6%
Been here since the Overcoming Bias days: 210, 14.0%
Referred by a friend: 199, 13.2%
Referred by a search engine: 114, 7.6%
Referred by other fiction: 17, 1.1%

[Amusing responses include "a rationalist that I follow on Tumblr", "I'm a student of tribal cultishness", and "It is difficult to recall details from the Before Time. Things were brighter, simpler, as in childhood or a dream. There has been much growth, change since then. But also loss. I can't remember where I found the link, is what I'm saying."]

Blog Referrals
Slate Star Codex: 40, 2.6%

Reddit: 25, 1.6%
Common Sense Atheism: 21, 1.3%
Hacker News: 20, 1.3%
Gwern: 13, 1.0%

VI. Other Categorical Data

Cryonics Status
Don't understand/never thought about it: 62, 4.1%
Don't want to: 361, 24.0%
Considering it: 551, 36.7%
Haven't gotten around to it: 272, 18.1%
Unavailable in my area: 126, 8.4%
Yes: 64, 4.3%

Type of Global Catastrophic Risk
Asteroid strike: 64, 4.3%
Economic/political collapse: 151, 10.0%
Environmental collapse: 218, 14.5%
Nanotech/grey goo: 47, 3.1%
Nuclear war: 239, 15.8%
Pandemic (bioengineered): 310, 20.6%
Pandemic (natural): 113. 7.5%
Unfriendly AI: 244, 16.2%

[Amusing answers include ennui/eaten by Internet, Friendly AI, "Greens so weaken the rich countries that barbarians conquer us", and Tumblr.]

Effective Altruism (do you self-identify)
Yes: 422, 28.1%
No: 758, 50.4%

[Despite some impressive outreach by the EA community, numbers are largely the same as last year]

Effective Altruism (do you participate in community)
Yes: 191, 12.7%
No: 987, 65.7%

Vegetarian
Vegan: 31, 2.1%
Vegetarian: 114, 7.6%
Other meat restriction: 252, 16.8%
Omnivore: 848, 56.4%

Paleo Diet
Yes: 33, 2.2%
Sometimes: 209, 13.9%
No: 1111, 73.9%

Food Substitutes
Most of my calories: 8. 0.5%
Sometimes: 101, 6.7%
Tried: 196, 13.0%
No: 1052, 70.0%

Gender Default
I only identify with my birth gender by default: 681, 45.3%
I strongly identify with my birth gender: 586, 39.0%

Books
<5: 198, 13.2%
5 - 10: 384, 25.5%
10 - 20: 328, 21.8%
20 - 50: 264, 17.6%
50 - 100: 105, 7.0%
> 100: 49, 3.3%

Birth Month
Jan: 109, 7.3%
Feb: 90, 6.0%
Mar: 123, 8.2%
Apr: 126, 8.4%
Jun: 107, 7.1%
Jul: 109, 7.3%
Aug: 120, 8.0%
Sep: 94, 6.3%

Oct: 111, 7.4%
Nov: 102, 6.8%
Dec: 106, 7.1%

[Despite my hope of something turning up here, these results don't deviate from chance]

Handedness
Right: 1170, 77.8%
Left: 143, 9.5%
Ambidextrous: 37, 2.5%
Unsure: 12, 0.8%

Previous Surveys
Yes: 757, 50.7%
No: 598, 39.8%

Favorite Less Wrong Posts (all > 5 listed)
An Alien God: 11
Joy In The Merely Real: 7
Dissolving Questions About Disease: 7
Politics Is The Mind Killer: 6
That Alien Message: 6
A Fable Of Science And Politics: 6
Belief In Belief: 5
Generalizing From One Example: 5
Schelling Fences On Slippery Slopes: 5
Tsuyoku Naritai: 5

VII. Numeric Data

Age: 27.67 + 8.679 (22, 26, 31) [1490]
IQ: 138.25 + 15.936 (130.25, 139, 146) [472]
SAT out of 1600: 1470.74 + 113.114 (1410, 1490, 1560) [395]
SAT out of 2400: 2210.75 + 188.94 (2140, 2250, 2320) [310]
ACT out of 36: 32.56 + 2.483 (31, 33, 35) [244]
Time in Community: 2010.97 + 2.174 (2010, 2011, 2013) [1317]
Time on LW: 15.73 + 95.75 (2, 5, 15) [1366]
Karma Score: 555.73 + 2181.791 (0, 0, 155) [1335]

P Many Worlds: 47.64 + 30.132 (20, 50, 75) [1261]
P Aliens: 71.52 + 34.364 (50, 90, 99) [1393]
P Aliens (Galaxy): 41.2 + 38.405 (2, 30, 80) [1379]
P Supernatural: 6.68 + 20.271 (0, 0, 1) [1386]
P God: 8.26 + 21.088 (0, 0.01, 3) [1376]
P Religion: 4.99 + 18.068 (0, 0, 0.5) [1384]
P Cryonics: 22.34 + 27.274 (2, 10, 30) [1399]
P Anti-Agathics: 24.63 + 29.569 (1, 10, 40) [1390]
P Simulation 24.31 + 28.2 (1, 10, 50) [1320]
P Warming 81.73 + 24.224 (80, 90, 98) [1394]
P Global Catastrophic Risk 72.14 + 25.620 (55, 80, 90) [1394]
Singularity: 2143.44 + 356.643 (2060, 2090, 2150) [1177]

[The mean for this question is almost entirely dependent on which stupid responses we choose to delete as outliers; the median practically never changes]

Abortion: 4.38 + 1.032 (4, 5, 5) [1341]
Immigration: 4 + 1.078 (3, 4, 5) [1310]
Taxes : 3.14 + 1.212 (2, 3, 4) [1410] (from 1 - should be lower to 5 - should be higher)
Minimum Wage: 3.21 + 1.359 (2, 3, 4) [1298] (from 1 - should be lower to 5 - should be higher)
Feminism: 3.67 + 1.221 (3, 4, 5) [1332]
Social Justice: 3.15 + 1.385 (2, 3, 4) [1309]
Human Biodiversity: 2.93 + 1.201 (2, 3, 4) [1321]
Basic Income: 3.94 + 1.087 (3, 4, 5) [1314]
Great Stagnation: 2.33 + .959 (2, 2, 3) [1302]
MIRI Mission: 3.90 + 1.062 (3, 4, 5) [1412]
MIRI Effectiveness: 3.23 + .897 (3, 3, 4) [1336]

[Remember, all of these are asking you to rate your belief in/agreement with the concept on a scale of 1 (bad) to 5 (great)]

Income: 54129.37 + 66818.904 (10,000, 30,800, 80,000) [923]
Charity: 1996.76 + 9492.71 (0, 100, 800) [1009]
MIRI/CFAR: 511.61 + 5516.608 (0, 0, 0) [1011]
XRisk: 62.50 + 575.260 (0, 0, 0) [980]
Older siblings: 0.51 + .914 (0, 0, 1) [1332]
Younger siblings: 1.08 + 1.127 (0, 1, 1) [1349]
Height: 178.06 + 11.767 (173, 179, 184) [1236]

Hours Online: 43.44 + 25.452 (25, 40, 60) [1221]
Bem Sex Role Masculinity: 42.54 + 9.670 (36, 42, 49) [1032]
Bem Sex Role Femininity: 42.68 + 9.754 (36, 43, 50) [1031]
Right Hand: .97 + 0.67 (.94, .97, 1.00)
Left Hand: .97 + .048 (.94, .97, 1.00)

VIII. Fishing Expeditions

[correlations, in descending order]

SAT Scores out of 1600/SAT Scores out of 2400 .844 (59)
P Supernatural/P God .697 (1365)
Feminism/Social Justice .671 (1299)
P God/P Religion .669 (1367)
P Supernatural/P Religion .631 (1372)
Charity Donations/MIRI and CFAR Donations .619 (985)
P Aliens/P Aliens 2 .607 (1376)
Taxes/Minimum Wage .587 (1287)
SAT Score out of 2400/ACT Score .575 (89)
Age/Number of Children .506 (1480)
P Cryonics/P Anti-Agathics .484 (1385)
SAT Score out of 1600/ACT Score .480 (81)
Minimum Wage/Social Justice .456 (1267)
Taxes/Social Justice .427 (1281)
Taxes/Feminism .414 (1299)
MIRI Mission/MIRI Effectiveness .395 (1331)
P Warming/Taxes .385 (1261)
Taxes/Basic Income .383 (1285)
Minimum Wage/Feminism .378 (1286)
P God/Abortion -.378 (1266)
Immigration/Feminism .365 (1296)
P Supernatural/Abortion -.362 (1276)
Feminism/Human Biodiversity -.360 (1306)
MIRI and CFAR Donations/Other XRisk Charity Donations .345 (973)
Social Justice/Human Biodiversity -.341 (1288)
P Religion/Abortion -.326 (1275)
P Warming/Minimum Wage .324 (1248)
Minimum Wage/Basic Income .312 (1276)
P Warming/Basic Income .306 (1260)
Immigration/Social Justice .294 (1278)
P Anti-Agathics/MIRI Mission .293 (1351)
P Warming/Feminism .285 (1281)
P Many Worlds/P Anti-Agathics .276 (1245)
Social Justice/Femininity .267 (990)
Minimum Wage/Human Biodiversity -.264 (1274)
Immigration/Human Biodiversity -.263 (1286)
P Many Worlds/MIRI Mission .263 (1233)
P Aliens/P Warming .262 (1365)
P Warming/Social Justice .257 (1262)
Taxes/Human Biodiversity -.252 (1291)
Social Justice/Basic Income .251 (1281)
Feminism/Femininity .250 (1003)
Older Siblings/Younger Siblings -.243 (1321)
Charity Donations/Other XRisk Charity Donations .240 (957
P Anti-Agathics/P Simulation .238 (1312)
Abortion/Minimum Wage .229 (1293)
Feminism/Basic Income .227 (1297)
Abortion/Feminism .226 (1321)
P Cryonics/MIRI Mission .223 (1360)
Immigration/Basic Income .208 (1279)
P Many Worlds/P Cryonics .202 (1251)
Number of Current Partners/Femininity: .202 (1029)
P Warming/Immigration .202 (1260)
P Warming/Abortion .201 (1289)
Abortion/Taxes .198 (1304)
Age/P Simulation .197 (1313)
Political Interest/Masculinity .194 (1011)
P Cryonics/MIRI Effectiveness .191 (1285)
Abortion/Social Justice .191 (1301)
P Simulation/MIRI Mission .188 (1290)
P Many Worlds/P Warming .188 (1240)
Age/Number of Current Partners .184 (1480)
P Anti-Agathics/MIRI Effectiveness .183 (1277)
P Many Worlds/P Simulation .181 (1211)
Abortion/Immigration .181 (1304)
Number of Current Partners/Number of Children .180 (1484)

P Cryonics/P Simulation .174 (1315)
P Global Catastrophic Risk/MIRI Mission -.174 (1359)
Minimum Wage/Femininity .171 (981)
Abortion/Basic Income .170 (1302)
Age/P Cryonics -.165 (1391)
Immigration/Taxes .165 (1293)
P Warming/Human Biodiversity -.163 (1271)
P Aliens 2/Warming .160 (1353)
Abortion/Younger Siblings -.155 (1292)
P Religion/Meditate .155 (1189)
Feminism/Masculinity -.155 (1004)
Immigration/Femininity .155 (988)
P Supernatural/Basic Income -.153 (1246)
P Supernatural/P Warming -.152 (1361)
Number of Current Partners/Karma Score .152 (1332)
P Many Worlds/MIRI Effectiveness .152 (1181)
Age/MIRI Mission -.150 (1404)
P Religion/P Warming -.150 (1358)
P Religion/Basic Income -.146 (1245)
P God/Basic Income -.146 (1237)
Human Biodiversity/Femininity -.145 (999)
P God/P Warming -.144 (1351)
Taxes/Femininity .142 (987)
Number of Children/Younger Siblings .138 (1343)
Number of Current Partners/Masculinity: .137 (1030)
P Many Worlds/P God -.137 (1232)
Age/Charity Donations .133 (1002)
P Anti-Agathics/P Global Catastrophic Risk -.132 (1373)
P Warming/Masculinity -.132 (992)
P Global Catastrophic Risk/MIRI and CFAR Donations -.132 (982)
P Supernatural/Singularity .131 (1148)
God/Taxes -.130 (1240)
Age/P Anti-Agathics -.128 (1382)
P Aliens/Taxes .127(1258)
Feminism/Great Stagnation -.127 (1287)
P Many Worlds/P Supernatural -.127 (1241)
P Aliens/Abortion .126 (1284)
P Anti-Agathics/Great Stagnation -.126 (1248)
P Anti-Agathics/P Warming .125 (1370)
Age/P Aliens .124 (1386)
P Aliens/Minimum Wage .124 (1245)
P Aliens/P Global Catastrophic Risk .122 (1363)
Age/MIRI Effectiveness -.122 (1328)
Age/P Supernatural .120 (1370)
P Supernatural/MIRI Mission -.119 (1345)
P Many Worlds/P Religion -.119 (1238)
P Religion/MIRI Mission -.118 (1344)
Political Interest/Social Justice .118 (1304)
P Anti-Agathics/MIRI and CFAR Donations .118 (976)
Human Biodiversity/Basic Income -.115 (1262)
P Many Worlds/Abortion .115 (1166)
Age/Karma Score .114 (1327)
P Aliens/Feminism .114 (1277)
P Many Worlds/P Global Catastrophic Risk -.114 (1243)
Political Interest/Femininity .113 (1010)
Number of Children/P Simulation -.112 (1317)
P Religion/Younger Siblings .112 (1275)
P Supernatural/Taxes -.112 (1248)
Age/Masculinity .112 (1027)
Political Interest/Taxes .111 (1305)
P God/P Simulation .110 (1296)
P Many Worlds/Basic Income .110 (1139)
P Supernatural/Younger Siblings .109 (1274)
P Simulation/Basic Income .109 (1195)
Age/P Aliens 2 .107 (1371)
MIRI Mission/Basic Income .107 (1279)
Age/Great Stagnation .107 (1295)
P Many Worlds/P Aliens .107 (1253)
Number of Current Partners/Social Justice .106 (1304)
Human Biodiversity/Great Stagnation .105 (1285)
Number of Children/Abortion -.104 (1337)
Number of Current Partners/P Cryonics -.102 (1396)
MIRI Mission/Abortion .102 (1305)
Immigration/Great Stagnation -.101 (1269)
Age/Political Interest .100 (1339)
P Global Catastrophic Risk/Political Interest .099 (1295)
P Aliens/P Religion -.099 (1357)

P God/MIRI Mission -.098 (1335)
P Aliens/P Simulation .098 (1308)
Number of Current Partners/Immigration .098 (1305)
P God/Political Interest .098 (1274)
P Warming/P Global Catastrophic Risk .096 (1377)

In addition to the Left/Right factor we had last year, this data seems to me to have an Agrees with the Sequences Factor--the same people tend to believe in many-worlds, cryo, atheism, simulationism, MIRI's mission and effectiveness, anti-agathics, etc. Weirdly, belief in global catastrophic risk is negatively correlated with most of the Agrees with Sequences things. Someone who actually knows how to do statistics should run a factor analysis on this data.

IX. Digit Ratios

After sanitizing the digit ratio numbers, the following correlations came up:

Digit ratio R hand was correlated with masculinity at a level of -0.180 $p < 0.01$
Digit ratio L hand was correlated with masculinity at a level of -0.181 $p < 0.01$
Digit ratio R hand was slightly correlated with femininity at a level of +0.116 $p < 0.05$

Holy #@!$ the feminism thing ACTUALLY HELD UP. There is a 0.144 correlation between right-handed digit ratio and feminism, $p < 0.01$. And an 0.112 correlation between left-handed digit ratio and feminism, $p < 0.05$.

The only other political position that correlates with digit ratio is immigration. There is a 0.138 correlation between left-handed digit ratio and believe in open borders $p < 0.01$, and an 0.111 correlation between right-handed digit ratio and belief in open borders, $p < 0.05$.

No digit correlation with abortion, taxes, minimum wage, social justice, human biodiversity, basic income, or great stagnation.

Okay, need to rule out that this is all confounded by gender. I ran a few analyses on men and women separately.

On men alone, the connection to masculinity holds up. Restricting sample size to men, left-handed digit ratio corresponds to masculinity with at -0.157, $p < 0.01$. Left handed at -0.134, $p < 0.05$. Right-handed correlates with femininity at 0.120, $p < 0.05$. The feminism correlation holds up. Restricting sample size to men, right-handed digit ratio correlates with feminism at a level of 0.149, $p < 0.01$. Left handed just barely fails to correlate. Both right and left correlate with immigration at 0.135, $p < 0.05$.

On women alone, the Bem masculinity correlation is the highest correlation we're going to get in this entire study. Right hand is -0.433, $p < 0.01$. Left hand is -0.299, $p < 0.05$. Femininity trends toward significance but doesn't get there. The feminism correlation trends toward significance but doesn't get there. In general there was too small a sample size of women to pick up anything but the most whopping effects.

Since digit ratio is related to testosterone and testosterone sometimes affects risk-taking, I wondered if it would correlate with any of the calibration answers. I selected people who had answered Calibration Question 5 incorrectly and ran an analysis to see if digit ratio was correlated with tendency to be more confident in the incorrect answer. No effect was found.

Other things that didn't correlate with digit ratio: IQ, SAT, number of current partners, tendency to work in mathematical professions.

...I still can't believe this actually worked. The finger-length/feminism connection ACTUALLY WORKED. What a world. What a world. Someone may want to double-check these results before I get *too* excited.

X. Calibration

There were ten calibration questions on this year's survey. Along with answers, they were:

1. What is the largest bone in the body? Femur
2. What state was President Obama born in? Hawaii
3. Off the coast of what country was the battle of Trafalgar fought? Spain
4. What Norse God was called the All-Father? Odin
5. Who won the 1936 Nobel Prize for his work in quantum physics? Heisenberg
6. Which planet has the highest density? Earth
7. Which Bible character was married to Rachel and Leah? Jacob
8. What organelle is called "the powerhouse of the cell"? Mitochondria
9. What country has the fourth-highest population? Indonesia
10. What is the best-selling computer game? Minecraft

I ran calibration scores for everybody based on how well they did on the ten calibration questions. These failed to correlate with IQ, SAT, LW karma, or any of the things you might expect to be measures of either intelligence or previous training in calibration; they didn't differ by gender, correlates of community membership, or any mental illness [deleted section about correlating with MWI and MIRI, this was an artifact].

Your answers looked like this:

The red line represents perfect calibration. Where answers dip below the line, it means you were overconfident; when they go above, it means you were underconfident.

It looks to me like everyone was horrendously underconfident on all the easy questions, and horrendously overconfident on all the hard questions. To give an example of how horrendous, people who were 50% sure of their answers to question 10 got it right only 13% of the time; people who were 100% sure only got it right 44% of the time. Obviously those numbers should be 50% and 100% respectively.

This builds upon results from previous surveys in which your calibration was also horrible. This is not a human universal - people who put even a small amount of training into calibration can become very well calibrated very quickly. This is a sign that most Less Wrongers continue to neglect the very basics of rationality and are incapable of judging how much evidence they have on a given issue. Veterans of the site do no better than newbies on this measure.

XI. Wrapping Up

To show my appreciation for everyone completing this survey, including the arduous digit ratio measurements, I have randomly chosen a person to receive a $30 monetary prize. That person is...the person using the public key "The World Is Quiet Here". If that person tells me their private key, I will give them $30.

I have removed 73 people who wished to remain private, deleted the Private Keys, and sanitized a very small amount of data. Aside from that, here are the raw survey results for your viewing and analyzing pleasure:

(as Excel)

(as SPSS)

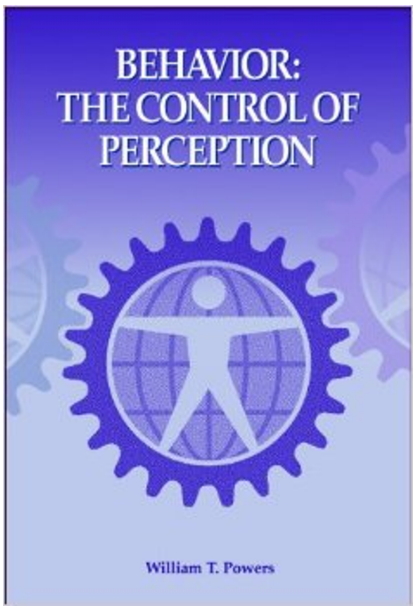
(as CSV)

Bill Gates: problem of strong AI with conflicting goals "very worthy of study and time"

> Steven Levy: Let me ask an unrelated question about the raging debate over whether artificial intelligence poses a threat to society, or even the survival of humanity. Where do you stand?
>
> Bill Gates: I think it's definitely important to worry about. There are two AI threats that are worth distinguishing. One is that AI does enough labor substitution fast enough to change work policies, or [affect] the creation of new jobs that humans are uniquely adapted to—the jobs that give you a sense of purpose and worth. We haven't run into that yet. I don't think it's a dramatic problem in the next ten years but if you take the next 20 to 30 it could be. Then there's the longer-term problem of so-called strong AI, where it controls resources, so its goals are somehow conflicting with the goals of human systems. Both of those things are very worthy of study and time. I am certainly not in the camp that believes we ought to stop things or slow things down because of that. But you can definitely put me more in the Elon Musk, Bill Joy camp than, let's say, the Google camp on that one.

"Bill Gates on Mobile Banking, Connecting the World and AI", Medium, 2015-01-21

Behavior: The Control of Perception



This is the second of three posts dealing with control theory and Behavior: The Control of Perception by William Powers. The previous post gave an introduction to control theory, in the hopes that a shared language will help communicate the models the book is discussing. This post discusses the model introduced in the book. The next post will provide commentary on the model and what I see as its implications, for both LW and AI.

B:CP was published in 1973 by William Powers, who was a controls engineer before he turned his attention to psychology. Perhaps unsurprisingly, he thought that the best lens for psychology was the one he had been trained in, and several sections of the book contrast his approach with the behaviorist approach. This debate is before my time, and so I find it mostly uninteresting, and will only bring it up when I think the contrast clarifies the difference in methodology, focusing instead on the meat of his model.

The first five chapters of B:CP introduce analog computers and feedback loops, and make the claim that it's better to model the nervous system as an analog computer, with continuous neural currents (with the strength of the current determined by the rate of the underlying impulses and number of branches, as each impulse has the same strength) rather than as a digital computer. On the macroscale, this seems unobjectionable, and while it makes the model clearer I'm not sure that it's necessary. He also steps through how to physically instantiate a number of useful mathematical functions with a handful of neurons; in general, I'll ignore that detailed treatment but you should trust that the book makes a much more detailed argument for the physical plausibility of this model than I do here.

The sixth chapter discusses the idea of hierarchical modeling. We saw a bit of that in the last post, with the example of a satellite that had two control systems, one which used sense data of the thruster impulses and the rotation to determine the inertia model, and the other which uses the rotation and the inertia model to determine the thruster impulses. The key point here is that the models are *inherently local*, and thus can be separated into units. The first model doesn't have to know that there's another feedback loop; it just puts the sense data it receives through a formula, and uses another formula to update its memory, which has the property of reducing the error of its model. Another way to look at this is that control systems are, in some sense, agnostic of what they're sensing and what they're doing, and their reference level comes from the environment similar to any other input. While one might see the two satellite systems as not being stacked, when one control circuit has no outputs except to vary the reference levels of other control circuits, it makes sense to see the reference-setting control circuit as superior in the hierarchical organization of the system.

There's a key insight hidden there which is probably best to show by example. The next five chapters of B:CP step through five levels in the hierarchy. Imagine this

section as building a model of a human as a robot- there are output devices (muscles and glands and so on) and input devices (sensory nerve endings) that are connected to the environment. Powers discusses both output and input, but here I'll just discuss output for brevity's sake.

Powers calls the first level *intensity*, and it deals directly with those output and input devices. Consider a muscle; the control loop there might have a reference tension in the muscle that it acts to maintain, and that tension corresponds to the outside environment. From the point of view of measured units, the control loops here convert between some physical quantity and neural current.

Powers calls the second level *sensation*, and it deals with combinations of first level sensors. As we've put all of the actual muscular effort of the arm and hand into the first level, the second level is one level of abstraction up. Powers suggests that the arm and hand have about 27 independent movements, and each movement represents some vector in the many-dimensional space of however many first order control loops there are. (Flexing the wrist, for example, might reflect an increase of effort intensity of muscles in the forearm on one side and a decrease of effort intensity of muscles on the other side.) Note that from this point on, the measured units are all the same--amperes of neural current--and that means we can combine unlike physical things because they have some conceptual similarity. This is the level of clustering where it starts to make sense to talk about a 'wrist,' or at least particular joints in it, or something like a 'lemonade taste,' which exists as a mental combination of the levels of sugar and acid in a liquid. The value of a hierarchy also starts to become clear- when we want to flex the wrist, we don't have to calculate what command to send to each individual muscle- we simply set a new reference level for the wrist-controller, and it adjusts the reference levels for many different muscle-controllers.

Powers calls the third level *configuration*, and it deals with combinations of second level loops. The example here would be the position of the joints in the arm or hand, such as positioning the hand to perform the Vulcan salute.

Powers calls the fourth level *transition*, and it deals with combinations of third level loops, as well as integration or differentiation of them. A third order control loop might put the mouth and vocal cords where they need to be to make a particular note, and a fourth order control loop could vary the note that third order control loop is trying to hit in order to create a pitch that rises at a certain rate.

Powers calls the fifth level *sequence*, and it deals with combinations and patterns of the fourth level loops. Here we see patterns of behavior- a higher level can direct a loop at this level to 'walk' to a particular place, and then the orders go down until muscles contract at the lowest level.

The key insight is that we can, whenever we identify a layer, see that layer as *part of the environment* of the hierarchy above that layer. At the 0th layer, we have the universe- at the 1st layer, we have a body, at the 2nd layer, we have a brain (and maybe a spine and other bits of the nervous system). Moving up layers in the organization is like peeling an onion; we consider smaller and smaller portions of the physical brain and consider more and more abstract concepts.

I'm not a neuroscientist, but I believe that until this point Powers's account would attract little controversy. The actual organization structure of the body is not as cleanly pyramidal as this brief summary makes it sound, but Powers acknowledges as much and the view remains broadly accurate and useful. There's ample neurological

evidence to support that there are parts of the brain that do the particular functions we would expect the various orders of control loops to do, and the interested reader should take a look at the book.

Where Powers's argument becomes more novel, speculative, and contentious is the that the levels keep going up, with the same basic architecture. Instead of an layered onion body wrapped around an opaque homunculus mind, it's an onion all the way to the center- which Powers speculates ends at around the 9th level. (More recent work, I believe, estimates closer to 11 levels.) The hierarchy isn't necessarily neat, with clearly identifiable levels, but there is some sort of hierarchical block diagram that stretches from terminal goal**s** to the environment. He identifies the levels as *relationships*, algorithms (which he calls *program control*), *principles*, and *system concepts*. As the abstractness of the concepts would suggest, their treatment is more vague and he manages to combine them all into a single chapter, with very little of the empirical justification that filled earlier chapters.

This seems inherently plausible to me:

1. It's parsimonious to use the same approach to signal processing everywhere, and it seems easier to just add on another layer of signal processing (which allows more than a linear increase in potential complexity of organism behavior) than to create an entirely new kind of brain structure.
2. Deep learning and similar approaches in machine learning can fit comparable architecture in an unsupervised fashion. My understanding of the crossover between machine learning and neuroscience is that we understand machine vision the best, and many good algorithms line up with what we see in human brains--pixels get aggregated to make edges which get aggregated to make shapes, and so on up the line. So we see the neural hierarchies this model predicts, but this isn't too much of a surprise because hierarchies are the easiest structures to detect and interpret.

What is meant by "terminal goals"? Well, control systems have to get their reference from somewhere, and the structure of the brain can't be "turtles all the way up." Eventually there should be a measured variable, like "hunger," which is compared to some reference, and any difference between the variable and the reference leads to action targeted at reducing the difference.

That reference could be genetic/instinctual, or determined by early experience, or modified by chemicals, or so on, but the point is that it isn't the feedback of a *neural* control loop above it. Chapter 14 discusses learning as the reorganization of the control system, and the processes used there seem potentially sufficient to explain where the reference levels and the terminal goals come from.

Indeed, the entire remainder of the book, discussing emotion, conflict, and so on, fleshes out this perspective in a full way that I could only begin to touch on here, so I will simply recommend reading the book if you're interested in his model. Here's a sample on conflict:

> Conflict is an encounter between two control systems, an encounter of a specific kind. In effect, the two control systems attempt to control the same quantity, but with respect to two different reference levels. For one system to correct an error, the other system must experience error. There is no way for both systems to experience zero error at the same time. Therefore the outputs of the system must act on the shared controlled quantity in opposite directions.
>
> If both systems are reasonably sensitive to error, and the two reference levels are far apart, there will be a range of values of the controlled quantity (between the reference levels) throughout which each system will contain an error signal so large that the output of each system will be solidly at its maximum. These two outputs, if about equal, will cancel, leaving essentially no net output to affect the controlled quantity. Certainly the net output cannot change as the “controlled” quantity changes in this region between the reference levels, since both outputs remain at maximum.
>
> This means there is a range of values over which the controlled quantity cannot be protected against disturbance any more. Any moderate disturbance will change the controlled quantity, and this will change the perceptual signals in the two control systems. As long as neither reference level is closely approached, there will be no reaction to these changes on the part of the conflicted systems.
>
> When a disturbance forces the controlled quantity close enough to either reference level, however, there *will* be a reaction. The control system experiencing lessened error will relax, unbalancing the net output in the direction of the *other* reference level. As a result, the conflicted pair of systems will act like a single system having a “virtual reference level,” between the two actual ones. A large dead zone will exist around the virtual reference level, within which there is little or no control.
>
> In terms of real behavior, this model of conflict seems to have the right properties. Consider a person who has two goals: one to be a nice guy, and the other to be a strong, self-sufficient person. If he perceives these two conditions in the “right” way (for conflict) he may find himself wanting to be deferential and pleasant, and at the same time wanting to speak up firmly for his rights. As a result, he does neither. He drifts in a state between, his attitude fluctuating with every change in external circumstances, undirected. When cajoled and coaxed enough he may find himself beginning to warm up, smile, and think of a pleasant remark, but immediately he realizes that he is being manipulated and resentfully breaks off communication or utters a cutting remark. On the other hand if circumstances lead him to begin defending himself against unfair treatment, his first strong words fill him with remorse and he blunts his defense with an apologetic giggle. He can react only when pushed to one extreme or the other, and his reaction takes him back to the uncontrolled middle ground.

So what was that about behaviorism?

According to Powers, most behaviorists thought in terms of 'stimulus->response,' where you could model a creature as a lookup table that would respond in a particular way to a particular stimulus. This has some obvious problems--how do we cluster stimuli? Someone saying "I love you" means very different things depending on the context. If the creature has a goal that depends on a relationship between entities, like wanting there to not be an unblocked line between their eyes and the sun, then you need to know the position of the sun to best model their response to any stimulus. Otherwise, if you just record what happens when you move a shade to the left, you'll notice that sometimes they move left and sometimes they move right. (Consider the difference between 1-place functions and 2-place functions.)

Powers discusses a particular experiment of neural stimulation in cats where the researchers couldn't easily interpret what some neurons were doing in behaviorist terms, because the cat would inconsistently move one way or another, but the control theory view parsimoniously explained the neurons as being higher order, which meant that the original position had to be taken into account to determine what the error was when the reference was adjusted by electrical stimulation, as it's the error that determines the response rather than just the reference.

If we want to have a lookup table in which the *entire life history* of the creature is the input, then figuring out what this table looks like is basically impossible. We want something that's complex enough to encode realistic behavior without being complex enough to encode unrealistic behavior--that is, we want the structure of our model to match the structure of the actual brain and behavior, and it looks like the control theory view is a strong candidate.

Unfortunately, I'm not an expert in this field, so I can't tell you what the state of the academic discussion looks like now. I get the impression that a number of psychologists have at least partly bought into the BCP paradigm (called Perceptual Control Theory) and have been working on their interests for decades, but it doesn't seem to have swept the field. As a general comment, controversies like this are often resolved by synthesis rather than the complete victory of one side over the other. If modern psychologists have learned a bit of the hierarchical control systems viewpoint and avoided the worst silliness of the past, then the historic criticisms are no longer appropriate and most of the low-hanging fruit from adopting this view haven already been picked.

Next: a comparison with utility, discussion of previous discussion on LW, and some thoughts on how thinking about control systems can impact thinking about AI.

Elon Musk donates $10M to the Future of Life Institute to keep AI beneficial

> We are delighted to report that technology inventor Elon Musk, creator of Tesla and SpaceX, has decided to donate $10M to the Future of Life Institute to run a global research program aimed at keeping AI beneficial to humanity.
>
> There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. A long list of leading AI-researchers have signed an open letter calling for research aimed at ensuring that AI systems are robust and beneficial, doing what we want them to do. Musk's donation aims to support precisely this type of research: "*Here are all these leading AI researchers saying that AI safety is important*", says Elon Musk. "*I agree with them, so I'm today committing $10M to support research aimed at keeping AI beneficial for humanity.*"
>
> [...] The $10M program will be administered by the Future of Life Institute, a non-profit organization whose scientific advisory board includes AI-researchers Stuart Russell and Francesca Rossi. [...]
>
> The research supported by the program will be carried out around the globe via an open grants competition, through an application portal at http://futureoflife.org that will open by Thursday January 22. The plan is to award the majority of the grant funds to AI researchers, and the remainder to AI-related research involving other fields such as economics, law, ethics and policy (a detailed list of examples can be found here [PDF]). "*Anybody can send in a grant proposal, and the best ideas will win regardless of whether they come from academia, industry or elsewhere*", says FLI co-founder Viktoriya Krakovna.
>
> [...] Along with research grants, the program will also include meetings and outreach programs aimed at bringing together academic AI researchers, industry AI developers and other key constituents to continue exploring how to maximize the societal benefits of AI; one such meeting was held in Puerto Rico last week with many of the open-letter signatories.

Elon Musk donates $10M to keep AI beneficial, Future of Life Institute, Thursday January 15, 2015

Immortality: A Practical Guide

Immortality: A Practical Guide

Introduction

This article is about how to increase one's own chances of living forever or, failing that, living for a long time. To be clear, this guide defines death as the long-term loss of one's consciousness and defines immortality as never-ending life. For those who would like less lengthy information on decreasing one's risk of death, I recommend reading the sections "Can we become immortal," "Should we try to become immortal," and "Cryonics," in this guide, along with the article *Lifestyle Interventions to Increase Longevity*.

This article does not discuss how to treat specific disease you may have. It is not intended as a substitute for the medical advice of physicians. You should consult a physician with respect to any symptoms that may require diagnosis or medical attention.

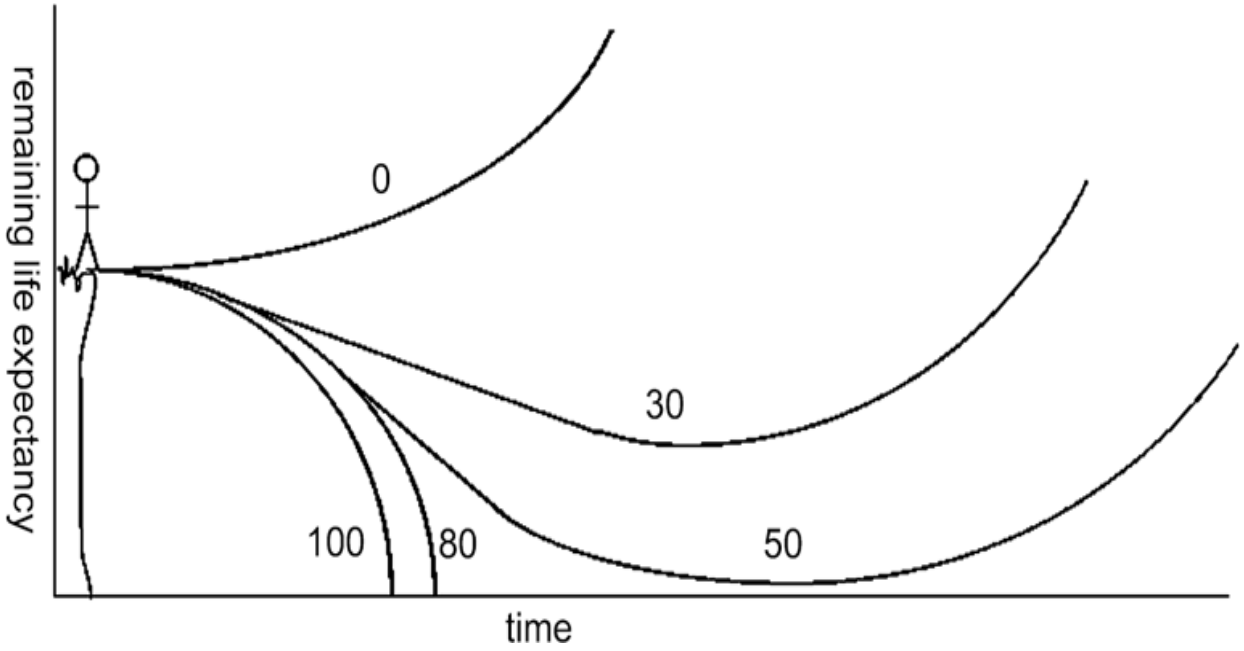
When reading about the effect sizes in scientific studies, keep in mind that many scientific studies report false-positives and are biased,[101] though I have tried to minimize this by maximizing the quality of the studies used. Meta-analyses and scientific reviews seem to typically be of higher quality than other study types, but are still subject to biases.[114]

Corrections, criticisms, and suggestions for new topics are greatly appreciated. I've tried to write this article tersely, so feedback on doing so would be especially appreciated. Apologies if the article's font type, size and color isn't standard on Less Wrong; I made it in google docs without being aware of Less Wrong's standard and it would take too much work changing the style of the entire article.

Contents

Can we become immortal?

In order to potentially live forever, one never needs to make it impossible to die; one instead just needs to have one's life expectancy increase faster than time passes, a concept known as the longevity escape velocity.[61] For example, if one had a 10% chance of dying in their first century of life, but their chance of death decreased by 90% at the end of each century, then one's chance of ever dying would be be $0.1 + 0.1^2 + 0.1^3\ldots = 0.11\ldots = 11.11...\%$. When applied to risk of death from aging, this akin to one's remaining life expectancy after jumping off a cliff while being affected by gravity and jet propulsion, with gravity being akin to aging and jet propulsion being akin to anti-aging (rejuvenation) therapies, as shown below.



The numbers in the above figure denote plausible ages of individuals when the first rejuvenation therapies arrive. A 30% increase in healthy lifespan would give the users of first-generation rejuvenation therapies 20 years to benefit from second-generation rejuvenation therapies, which could give an additional 30% increase if life span, ad infinitum.[61]

As for causes of death, many deaths are strongly age-related. The proportion of deaths that are caused by aging in the industrial world approaches *90%*.[53] Thus, I suppose postponing aging would drastically increase life expectancy.

As for efforts against aging, the SENS Research foundation and Science for Life Extension are charitable foundations for trying to cure aging.[54, 55] Additionally, Calico, a Google-backed company, and AbbVie, a large pharmaceutical company, have each committed fund $250 million to cure aging.[56]

I speculate that one could additionally decrease risk of death by becoming a cyborg, as mechanical bodies seem easier to maintain than biological ones, though I've found no articles discussing this.

Similar to becoming a cyborg, another potential method of decreasing one's risk of death is mind uploading, which is, roughly speaking, the transfer of most or all of one's mental contents into a computer.[62] However, there are some concerns about the transfer creating a copy of one's consciousness, rather than being the same consciousness. This issue is made very apparent if the mind-uploaded process leaves the original mind intact, making it seem unlikely that one's consciousness was transferred to the new body.[63] Eliezer Yudkowsky doesn't seem to believe this is an issue, though I haven't found a citation for this.

With regard to consciousness, it seems that most individuals believe that the consciousness in one's body is the "same" consciousness as the one that was in one's body in the past and will be in it in the future. However, I know of no evidence for this. If one's consciousness isn't the same of the one in one's body in the future, and one defined death as one's consciousness permanently ending, then I suppose one can't prevent death for any time at all. Surprisingly, I've found no articles discussing this possibility.

Although curing aging, becoming a cyborg, and mind uploading may prevent death from disease, they still seem to leave oneself vulnerable to accidents, murder, suicide, and existential catastrophes. I speculate that these problems could be solved by giving an artificial superintelligence the ability to take control of one's body in order to prevent such deaths from occurring. Of course, this possibility is currently unavailable.

Another potential cause of death is the Sun expanding, which could render Earth uninhabitable in roughly one billion years. Death from this could be prevented by colonizing other planets in the solar system, although eventually the sun would render the rest of the solar system uninhabitable. After this, one could potentially inhabit other stars; it is expected that stars will remain for roughly 10 quintillion years, although some theories predict that the universe will be destroyed in a mere 20 billion years. To continue surviving, one could potentially go to other universes.[64] Additionally, there are ideas for space-time crystals that could process information even after heat death (i.e. the "end of the universe"),[65] so perhaps one could make oneself composed of the space-time crystals via mind uploading or another technique. There could also be other methods of surviving the conventional end of the universe, and life could potentially have 10 quintillion years to find them.

Yet another potential cause of death is living in a computer simulation that is ended. The probability of one living in a computer simulation actually seems to not be very improbable. Nick Bostrom argues that:

> ...*at least one* of the following propositions is true: (1) The fraction of human-level civilizations that reach a posthuman stage is very close to zero; (2) The fraction of posthuman civilizations that are interested in running ancestor-simulations is very close to zero; (3) The fraction of all people with our kind of experiences that are living in a simulation is very close to one.

The argument for this is here.[100]

If one does die, one could potentially be revived. Cryonics, discussed later in this article, may help in this. Additionally, I suppose one could possibly be revived if future intelligences continually create new conscious individuals and eventually create one of them that have one's "own" consciousness, though consciousness remains a mystery, so this may not be plausible, and I've found no articles discussing this possibility. If the probability of one's consciousness being revived per unit time does not approach or equal zero as time approaches infinity, then I suppose one is bound to become conscious again, though this scenario may be unlikely. Again, I've found no articles discussing this possibility.

As already discussed, in order to be live forever, one must either be revived after dying or prevent death from the consciousness in one's body not being the same as the one that will be in one's body in the future, accidents, aging, the sun dying, the universe dying, being in a simulation and having it end, and other, unknown, causes. Keep in mind that adding extra details that aren't guaranteed to be true can only make events less probable, and that people often don't account for this.[66] A spreadsheet for estimating one's chance of living forever is here.

Should we try to become immortal?

Before deciding whether one should try to become immortal, I suggest learning about the cognitive biases scope insensitivity, hyperbolic discounting, and bias blind spot if you don't know currently know about them. Also, keep in mind that one study found that simply informing people of a cognitive bias made them no less likely to fall prey to it. A study also found that people only partially adjusted for cognitive biases after being told that informing people of a cognitive bias made them no less likely to fall prey to it.[67]

Many articles arguing against immortality are found via a quick google search, including this, this, this, and this. This article along with its comments discusses counter-arguments to many of these arguments. *The Fable of the Dragon Tyrant* provides an argument for curing aging, which can be extended to be an argument against mortality as a whole. I suggest reading it.

One can also evaluate the utility of immortality via decision theory. Assuming individuals receive a finite amount of utility per unit time such that it is never less than some above-zero constant, living forever would give infinitely more utility than living for a finite amount of time. Using these assumptions, in order to maximize utility, one should be willing to accept any finite cost to become immortal. However, the situation is complicated when one considers the potential of becoming immortal and receiving an infinite positive utility unintentionally, in which case one would receive infinite expected utility regardless of if one tried to become immortal. Additionally, if

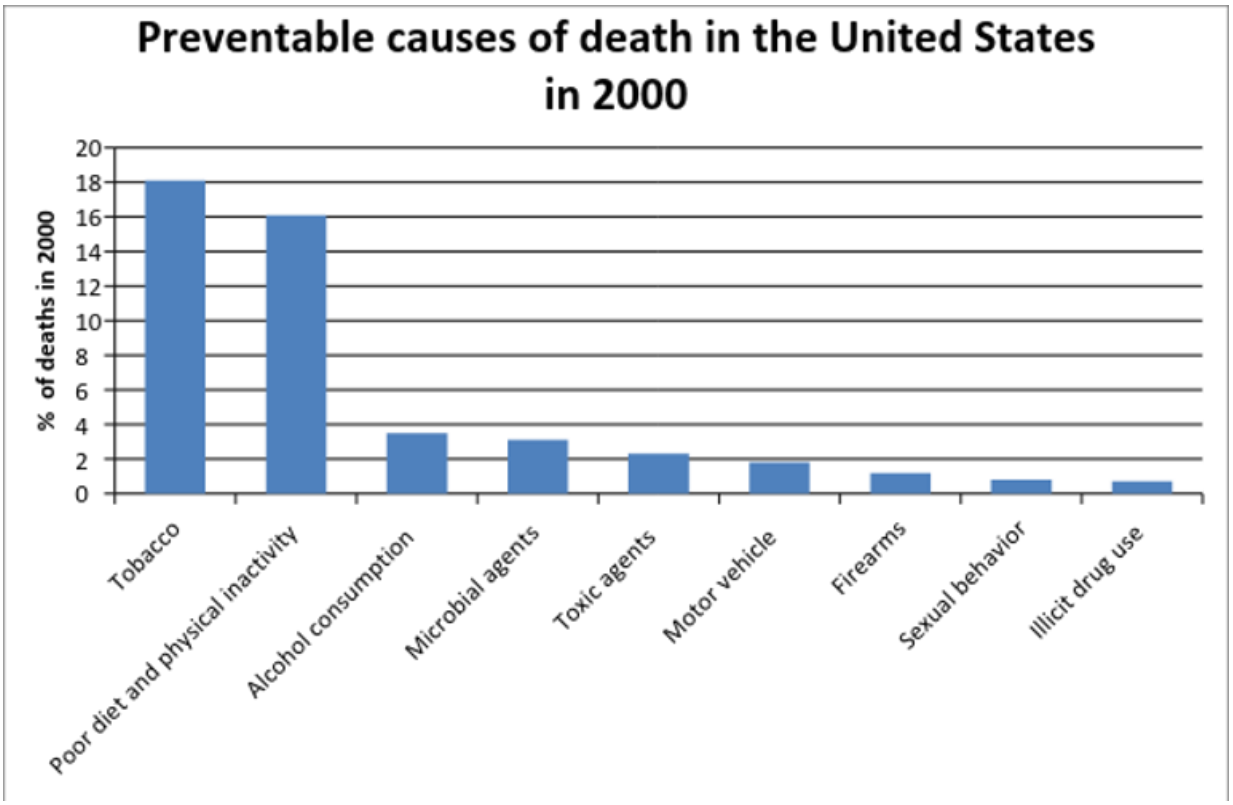
one both has the chance of receiving infinitely high and infinitely low utility, one's expected utility would be undefined. Infinite utilities are discussed in "Infinite Ethics" by Nick Bostrom.

For those interested in decreasing existential risk, living for a very long time, albeit not necessarily forever, may give one more opportunity to do so. This idea can be generalized to many goals one has in life.

On whether one can influence one's chances of becoming immortal, studies have shown that only roughly 20-30% of longevity in humans is accounted for by genetic factors.[68] There are multiple actions one can to increase one's chances of living forever; these are what the rest of this article is about. Keep in mind that you should consider continuing reading this article even if you don't want to try to become immortal, as the article provides information on living longer, even if not forever, as well.

Relative importance of the different topics

The figure below gives the relative frequencies of preventable causes of death.



1

Some causes of death are excluded from the graph, but are still large causes of death. Most notably, 440,000 deaths in the US, *roughly one sixth of total deaths in the US* are estimated to be from preventable medical errors in hospitals.[2]

Risk calculators for cardiovascular disease are here and here. Though they seem very simplistic, they may be worth looking at and can probably be completed quickly.

Here are the frequencies of causes of deaths in the US in year 2010 based off of another classification:

- Heart disease: 596,577
- Cancer: 576,691
- Chronic lower respiratory diseases: 142,943
- Stroke (cerebrovascular diseases): 128,932
- Accidents (unintentional injuries): 126,438
- Alzheimer's disease: 84,974
- Diabetes: 73,831
- Influenza and Pneumonia: 53,826
- Nephritis, nephrotic syndrome, and nephrosis: 45,591

- Intentional self-harm (suicide): 39,518

113

Food

What to eat and drink

Keep in mind that the relationship between health and the consumption of types of substances aren't necessarily linear. I.e. some substances are beneficial in small amounts but harmful in large amounts, while others are beneficial in both small and large amounts, but consuming large amounts is no more beneficial than consuming small amounts.

Recommendations from The Nutrition Source

The Nutrition Source is part of the Harvard School of Public Health.

Its recommendations:

- Make ½ of your "plate" consist of a variety of fruits and a variety of vegetables, excluding potatoes, due to potatoes' negative effect on blood sugar. The Harvard School of Public Health doesn't seem to specify if this is based on calories or volume. It also doesn't explain what it means by plate, but presumably ½ of one's plate means ½ solid food consumed.
- Make ¼ of your plate consist of whole grains.
- Make ¼ of your plate consist of high-protein foods.
- Limit red meat consumption.
- Avoid processed meats.
- Consume monounsaturated and polyunsaturated fats in moderation; they are healthy.
- Avoid partially hydrogenated oils, which contain trans fats, which are unhealthy.
- Limit milk and dairy products to one to two servings per day.
- Limit juice to one small glass per day.
- It is important to eat seafood one or two times per week, particularly fatty (dark meat) fish that are richer in EPA and DHA.
- Limit diet drink consumption or consume in moderation.
- Avoid sugary drinks like soda, sports drinks, and energy drinks.[3]

Fat

The bottom line is that saturated fats and especially trans fats are unhealthy, while unsaturated fats are healthy and the types of unsaturated fats omega-3 and omega-6 fatty acids fats are essential. The proportion of calories from fat in one's diet isn't really linked with disease.

Saturated fat is unhealthy. It's generally a good idea to minimize saturated fat consumption. The latest Dietary Guidelines for Americans recommends consuming no more than 10% of calories from saturated fat, but the American Heart Association recommends consuming no more than 7% of calories from saturated fat. However, don't decrease nut, oil, and fish consumption to minimize saturated fat consumption. Foods that contain large amounts of saturated fat include red meat, butter, cheese, and ice cream.

Trans fats are especially unhealthy. For every 2% increase of calories from trans-fat, risk of coronary heart disease increases by 23%. The Federal Institute for Medicine states that there are no known requirements for trans fats for bodily functions, so their consumption should be minimized. Partially hydrogenated oils contain trans fats, and foods that contain trans fats are often processed foods. In the US, products can claim to have zero grams of trans fat if they have no more than 0.5 grams of trans fat. Products with no more than 0.5 grams of trans fat that still have non-negligible amounts of trans fat will probably have the ingredients "partially hydrogenated vegetable oils" or "vegetable shortening" in their ingredient list.

Unsaturated fats have beneficial effects, including improving cholesterol levels, easing inflammation, and stabilizing heart rhythms. The American Heart Association has set 8-10% of calories as a target for polyunsaturated fat consumption, though eating more polyunsaturated fat, around 15%of daily calories, in place of saturated fat may further lower heart disease risk. Consuming unsaturated fats instead of saturated fat also prevents insulin resistance, a precursor to diabetes. Monounsaturated fats and polyunsaturated fats are types of unsaturated fats.

Omega-3 fatty acids (omega-3 fats) are a type of unsaturated fat. There are two main types: Marine omega-3s and alpha-linolenic acid (ALA). Omega-3 fatty acids, especially marine omega-3s, are healthy. Though one can make most needed types of fats from other fats or substances consumed, omega-3 fat is an essential fat, meaning it is an important type of fat and cannot be made in the body, so they must come from food. Most americans don't get enough omega-3 fats.

Marine omega-3s are primarily found in fish, especially fatty (dark mean) fish. A comprehensive review found that eating roughly two grams per week of omega-3s from fish, equal to about one or two servings of fatty fish per week, decreased risk of death from heart disease by more than one-third. Though fish contain mercury, this is insignificant the positive health effects of their consumption (for the consumer, not the fish). However, it does benefit one's health to consult local advisories to determine how much local freshwater fish to consume.

ALA may be an essential nutrient, and increased ALA consumption may be beneficial. ALA is found in vegetable oils, nuts (especially walnuts), flax seeds, flaxseed oil, leafy vegetables, and some animal fat, especially those from grass-fed animals. ALA is primarily used as energy, but a very small amount of it is converted into marine omega-3s. ALA is the most common omega-3 in western diets.

Most Americans consume much more omega-6 fatty acids (omega-6 fats) than omega-3 fats. Omega-6 fat is an essential nutrient and its consumption is healthy. Some sources of it include corn and soybean oils. The Nutrition Sources stated that the theory that omega-3 fats are healthier than omega-6 fats isn't supported by evidence. However, in an image from the Nutrition Source, seafood omega-6 fats were ranked as healthier than plant omega-6 fats, which were ranked as healthier than monounsaturated fats, although such a ranking was to the best of my knowledge never stated in the text.[3]

Carbohydrates

There seems to be two main determinants of carbohydrate sources' effects on health: nutrition content and effect on blood sugar. The bottom line is that consuming whole grains and other less processed grains and decreasing refined grain consumption improves health. Additionally, moderately low carbohydrate diets can increase heart health as long as protein and fat comes from health sources, though the type of carbohydrate at least as important as the amount of carbohydrates in a diet.

Glycemic index and is a measure of how much food increases blood sugar levels. Consuming carbohydrates that cause blood-sugar spikes can increase risk of heart disease and diabetes at least as much as consuming too much saturated fat does. Some factors that increase the glycemic index of foods include:

- Being a refined grain as opposed to a whole grain.
- Being finely ground, which is why consuming whole grains in their whole form, such as rice, can be healthier than consuming them as bread.
- Having less fiber.
- Being more ripe, in the case of fruits and vegetables.
- Having a lower fat content, as meals with fat are converted more slowly into sugar.

Vegetables (excluding potatoes), fruits, whole grains, and beans, are healthier than other carbohydrates. Potatoes have a negative effect on blood sugar, due to their high glycemic index. Information on glycemic index and the index of various foods is here.

Whole grains also contain essential minerals such as magnesium, selenium, and copper, which may protect against some cancers. Refining grains takes away 50% of the grains' B vitamins, 90% of vitamin E, and virtually all fiber. Sugary drinks usually have little nutritional value.

Identifying whole grains as food that has at least one gram of fiber for every gram of carbohydrate is a more effective measure of healthfulness than identifying a whole grain as the first ingredient, any whole grain as the first ingredient without added sugars in the first 3 ingredients, the word "whole" before any grain ingredient, and the whole grain stamp.[3]

Protein

Proteins are broken down to form amino acids, which are needed for health. Though the body can make some amino acids by modifying others, some must come from food, which are called essential amino acids. The institute of medicine recommends that adults get a minimum of 0.8 grams of protein per kilogram of body weight per day, and sets the range of acceptable protein intake to 10-35% of calories per day. The Institute of Medicine recommends getting 10-35% of calories from protein each day. The US recommended daily allowance for protein is 46 grams per day for women over 18 and 56 grams per day for men over 18.

Animal products tend to give all essential amino acids, but other sources lack some essential amino acids. Thus, vegetarians need to consume a variety of sources of amino acids each day to get all needed types. Fish, chicken, beans, and nuts are healthy protein sources.[3]

Fiber

There are two types of fiber: soluble fiber and insoluble fiber. Both have important health benefits, so one should eat a variety of foods to get both.[94] The best sources of fiber are whole grains, fresh fruits and vegetables, legumes, and nuts.[3]

Micronutrients

There are many micronutrients in food; getting enough of them is important. Most healthy individuals can get sufficient micronutrients by consuming a wide variety of healthy foods, such as fruits, vegetables, whole grains, legumes, and lean meats and fish. However, supplementation may be necessary for some. Information about supplements is here.[110]

Concerning supplementation, potassium, iodine, and lithium supplementation are recommended in the first-place entry in the Quantified Health Prize, a contest on determining good mineral intake levels. However, others suggest that potassium supplementation isn't necessarily beneficial, as shown here. I'm somewhat skeptical that the supplements are beneficial, as I have not found other sources recommending their supplementation. The suggested supplementation levels are in the entry.

Note that food processing typically decreases micronutrient levels, as described here. In general, it seems cooking, draining and drying foods sizably, taking potentially half of nutrients away, while freezing and reheating take away relatively few nutrients.[111]

One micronutrient worth discussing is sodium. Some sodium is needed for health, but most Americans consume more sodium than needed. However, recommendations on ideal sodium levels vary. The US government recommends limiting sodium consumption to 2,300mg/day (one teaspoon). The American Heart Association recommends limiting sodium consumption to 1,500mg/day (⅔ of a teaspoon), especially for those who are over 50, have high or elevated blood pressure, have diabetes, or are African Americans[3] However, As RomeoStevens pointed out, the Institute of Medicine found that there's inconclusive evidence that decreasing sodium consumption below 2,300mg/day effects mortality,[115] and some meta-analyses have suggested that there is a U-shaped relationship between sodium and mortality.[116, 117]

Vitamin D is another micronutrient that's important for health. It can be obtained from food or made in the body after sun exposure. Most people who live farther north than San Francisco or don't go outside at least fifteen minutes when it's sunny are vitamin D deficient. Vitamin D deficiency is increases the risk of many chronic diseases including heart disease, infectious diseases, and some cancers. However, there is controversy about optimal vitamin D intake. The Institute of medicine recommends getting 600 to 4000 IU/day, though it acknowledged that there was no good evidence of harm at 4000 IU/day. The Nutrition Sources states that these recommendations are too low and fail to account for new evidence. The nutrition source states that for most people, supplements are the best source of vitamin D, but most multivitamins have too little vitamin D in them. The Nutrition Source recommends considering and talking to a doctor about taking an additional multivitamin if the you take less than 1000 IU of vitamin D and especially if you have little sun exposure.[3]

Blood pressure

Information on blood pressure is here in the section titled "Blood Pressure."

Cholesterol and triglycerides

Information on optimal amounts of cholesterol and triglycerides are here.

The biggest influences on cholesterol are fats and carbohydrates in one's diet, and cholesterol consumption generally has a far weaker influence. However, some people's cholesterol levels rise and fall very quickly with the amount of cholesterol consumed. For them, decreasing cholesterol consumption from food can have a considerable effect on cholesterol levels. Trial and error is currently the only way of determining if one's cholesterol levels risk and fall very quickly with the amount of cholesterol consumed.

Antioxidants

Despite their initial hype, randomized controlled trials have offered little support for the benefit is single antioxidants, though studies are inconclusive.[3]

Dietary reference intakes

For the numerically inclined, the Dietary Reference Intake provides quantitative guidelines on good nutrient consumption amounts for many nutrients, though it may be harder to use for some, due to its quantitative nature.

Drinks

The Nutrition Source and SFGate state that water is the best drink,[3, 112] though I don't know why it's considered healthier than drinks such as tea.

Unsweetened tea decreases the risk of many diseases, likely largely due to polyphenols, and antioxidant, in it. Despite antioxidants typically having little evidence of benefit, I suppose polyphenols are relatively beneficial. All teas have roughly the same levels of polyphenols except decaffeinated tea,[3] which has fewer polyphenols.[96] Research suggests that proteins and possibly fat in milk decrease the antioxidant capacity of tea.

It's considered safe to drink up to six cups of coffee per day. *Unsweetened* coffee is healthy and may decrease some disease risks, though coffee may slightly increase blood pressure. Some people may want to consider avoiding coffee or switching to decaf, especially women who are pregnant or people who have a hard time controlling their blood pressure or blood sugar. The nutrition source states that it's best to brew coffee with a paper filter to remove a substance that increases LDL cholesterol, despite consumed cholesterol typically having a very small effect on the body's cholesterol level.

Alcohol increases risk of diseases for some people[3] and decreases it for others.[3, 119] Heavy alcohol consumption is a major cause of preventable death in most countries. For some groups of people, especially pregnant people, people recovering from alcohol addiction, and people with liver disease, alcohol causes greater health risks and should be avoided. The likelihood of becoming addicted to alcohol can be genetically determined. Moderate drinking, generally defined as no more than one or two drinks per day for men, can increase colon and breast cancer risk, but these effects are offset by decreased heart disease and diabetes risk, especially in middle age, where heart disease begins to account for an increasingly large proportion of deaths. However, alcohol consumption won't decrease cardiovascular disease risk much for those who are thin, physically active, don't smoke, eat a healthy diet, and have no family history of heart disease. Some research suggests that red wine, particularly when consumed after a meal, has more cardiovascular benefits than beers or spirits, but alcohol choice has still little effect on disease risk. In one study, moderate drinkers were 30-35% less likely to have heart attacks than non-drinkers and men who drank daily had lower heart attack risk than those who drank once or twice per week.

There's no need to drink more than one or two glasses of milk per day. Less milk is fine if calcium is obtained from other sources.

The health effects of artificially sweetened drinks are largely unknown. Oddly, they may also cause weight gain. It's best to limit consuming them if one drinks them at all.

Sugary drinks can cause weight gain, as they aren't as filling as solid food and have high sugar. They also increase the risk of diabetes, heart disease, and other diseases. Fruit juice has more calories and less fiber than whole fruit and is reportedly no better than soft drinks.[3]

Solid food

Fruits and vegetables are an important part of a healthy diet. Eating a variety of them is as important as eating many of them.[3] Fish and nut consumption is also very healthy.[98]

Processed meat, on the other hand, is shockingly bad.[98] A meta-analysis found that processed meat consumption is associated with a *42% increased risk of coronary heart disease* (relative risk per 50g serving per day; 95% confidence interval: 1.07 - 1.89) and 19% increased risk of diabetes.[97] Despite this, a bit of red meat consumption has been found to be beneficial.[98] Consumption of well-done, fried, or barbecued meat has been associated with certain cancers, presumably due to carcinogens made in the meat from being cooked, though this link isn't definitive. The amount of carcinogens increases with increased cooking temperature (especially above 300ºF, increased cooking time, charring, or being exposed to smoke.[99]

Eating less than one egg per day doesn't increase heart disease risk in healthy individuals and can be part of a healthy diet.[3]

Organic foods have lower levels of pesticides than inorganic foods, though the residues of most organic and inorganic products don't exceed government safety threshold. Washing fresh fruits and vegetables in recommended, as it removes bacteria and some, though not all, pesticide residues. Organic foods probably aren't more nutritious than non-organic foods.[103]

When to eat and drink

A randomized controlled trial found an increase in blood sugar variation for subjects who skipped breakfast.[6] Increasing meal frequency and decreasing meal size appears to have some metabolic advantages, and doesn't appear to have metabolic disadvantages.[7] Note: old source; made in 1994 However, Mayo Clinic states that fasting for 1-2 days per week may increase heart health.[32] Perhaps it is optimal for health to fast, but to have high meal frequency when not fasting.

How much to eat

One's weight gain is directly proportional to the number of calories consumed divided by the number of calories burnt. Centers for Disease Control and Prevention (CDC) has guidelines for healthy weights and information on how to lose weight.

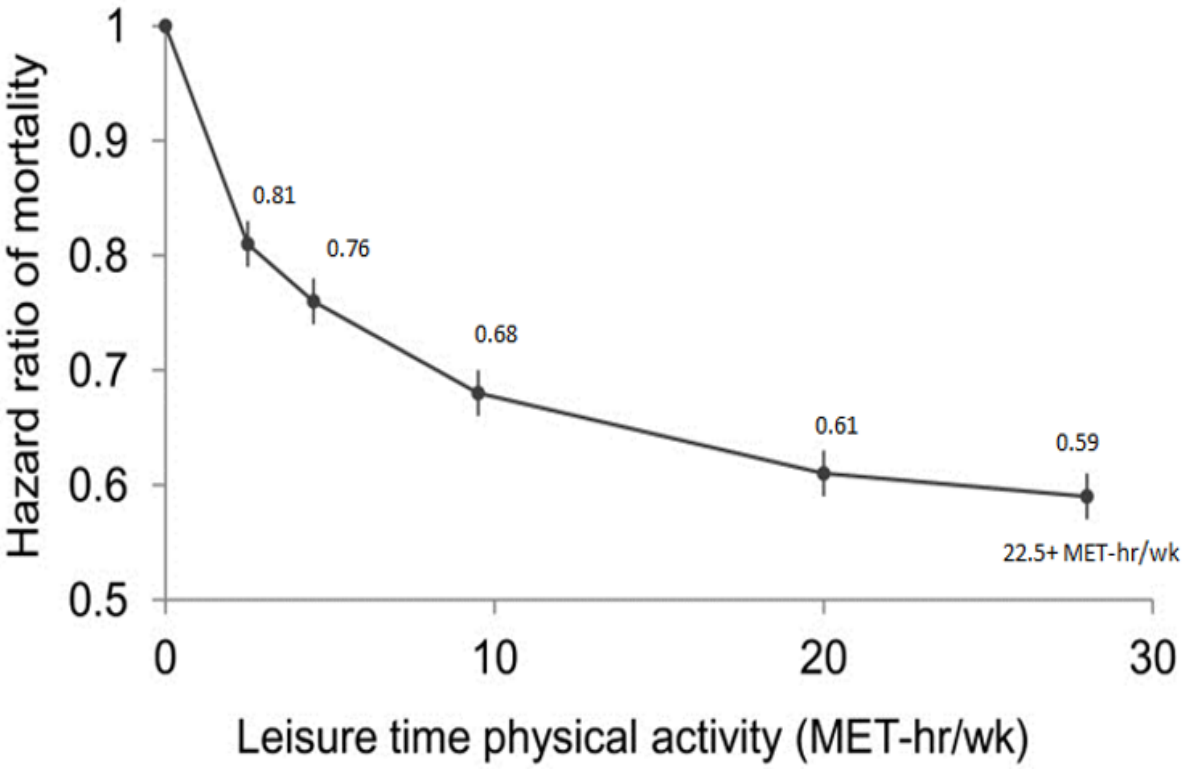
Some advocate restricting weight to a greater extent, which is known as calorie restriction. It's unknown whether calorie restriction increases lifespan in humans or not, but moderate calorie restriction with adequate nutrition decreases risk of obesity, type 2 diabetes, inflammation, hypertension, cardiovascular disease, and metabolic risk factors associated with cancer, and is the most effective way of consistently increasing lifespan in a variety of organisms. The CR Society has information on getting started on calorie restriction.[4]

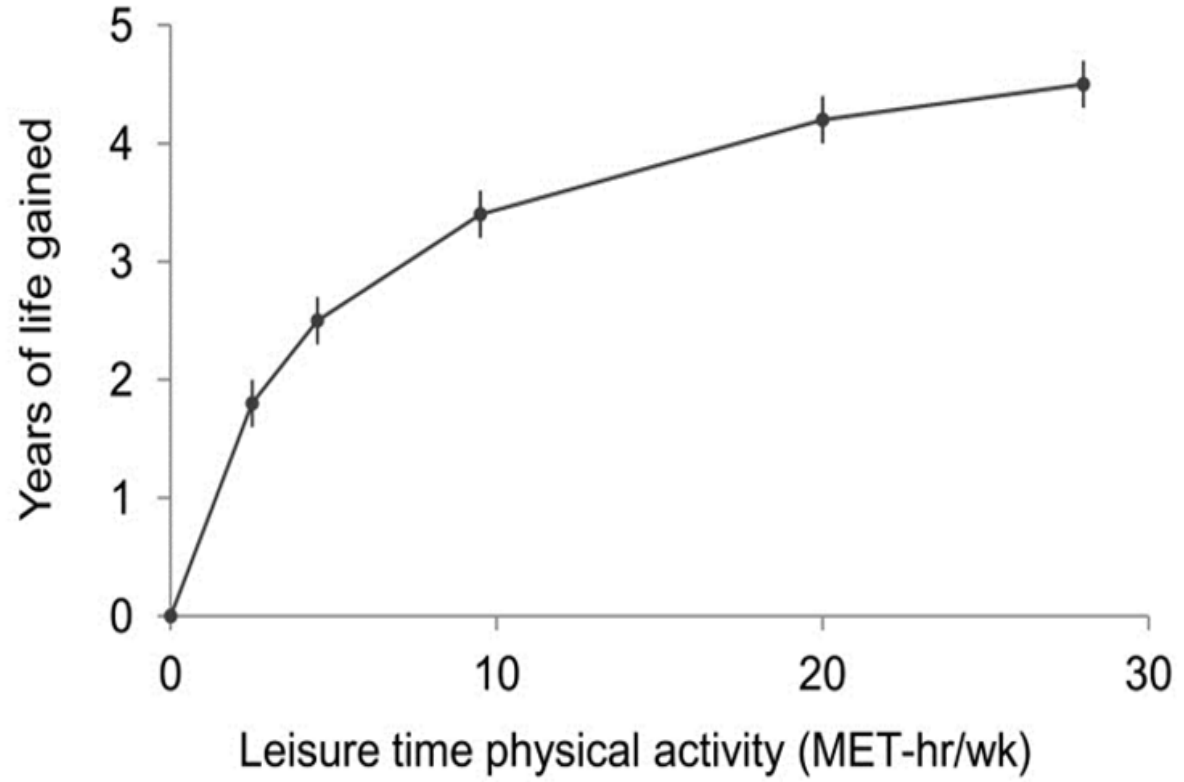
How much to drink

Generally, drinking enough to rarely feel thirsty and to have colorless or light yellow urine is usually sufficient. It's also possible to drink *too* much water. In general, drinking too much water is rare in healthy adults who eat an average American diet, although endurance athletes are at a higher risk.[10]

Exercise

A meta-analysis found the data in the following graphs for people aged over 40.


Hazard ratio of mortality
1
0.9
0.8
0.7
0.6
0.5
0.81
0.76
0.68
0.61
0.59
22.5+ MET-hr/wk
0
10
20
30
Leisure time physical activity (MET-hr/wk)


Years of life gained
5
4
3
2
1
0
0
10
20
30
Leisure time physical activity (MET-hr/wk)

A weekly total of roughly five hours of vigorous exercise has been identified by several studies to be the safe upper limit for life expectancy. It may be beneficial to take one or two days off from vigorous exercise per week and to limit chronic vigorous exercise to <= 60 min/day.[9] Based on the above, I my best guess for the optimal amount of exercise for longevity is roughly 30 MET-hr/wk. Calisthenics burn 6-10 METs/hr[11], so an example exercise routine to get this amount of exercise is doing calisthenics 38 minutes per day and 6 days/wk. Guides on how to exercise are available, e.g. this one.

Carcinogens

Carcinogens are cancer-causing substances. Since cancer causes death, decreasing exposure to carcinogens presumably decreases one's risk of death. Some foods are also carcinogenic, as discussed in the "Food" section.

Chemicals

Tobacco use is the greatest avoidable risk factor for cancer worldwide, causing roughly 22% of cancer deaths. Additionally, second hand smoke has been proven to cause lung cancer in nonsmoking adults.

Alcohol use is a risk factor for many types of cancer. The risk of cancer increases with the amount of alcohol consumed, and substantially increases if one is also a heavy smoker. The attributable fraction of cancer from alcohol use varies depending on gender, due to differences in consumption level. E.g. 22% of mouth and oropharynx cancer is attributable to cancer in men but only 9% is attributable to alcohol in women.

Environmental air pollution accounts for 1-4% of cancer.[84] Diesel exhaust is one type of carcinogenic air pollution. Those with the highest exposure to diesel exhaust are exposed to it occupationally. As for residential exposure, diesel exhaust is highest in homes near roads where traffic is heaviest. Limiting time spent near large sources of diesel exhaust decreases exposure. Benzene, another carcinogen, is found in gasoline and vehicle exhaust but exposure to it can also be cause by being in areas with unventilated fumes from gasoline, glues, solvents, paints, and art supplies. It can cause exposure from inhalation or skin contact.[86]

Some occupations exposure workers to occupational carcinogens.[84] A list of some of the occupations is here, all of which involve manual labor, except for hospital-related jobs.[87]

Infections

Infections are responsible for 6% of cancer deaths in developed nations.[84] Many of the infections are spread via sexual contact and sharing needles and some can be vaccinated against.[85]

Radiation

Ionizing radiation is carcinogenic to humans. Residential exposure to radon gas is estimated to cause 3-14% of lung cancers, which is the largest source of radon exposure for most people [84] Being exposed to radon and cigarette smoke together increases one's cancer risk much more than they do separately. There is much variation radon levels depending on where one lives and and radon is usually higher inside buildings, especially levels closer to the ground, such as basements. The EPA recommends taking action to reduce radon levels if they are greater than or equal to 4.0 pCi/L. Radon levels can be reduced by a qualified contractor. Reducing radon levels without proper training and equipment can increase instead of decrease them.[88]

Some medical tests can also increase exposure to radiation. The EPA estimates that exposure to 10 mSv from a medical imaging test increases risk of cancer by roughly 0.05%. To decrease exposure to radiation from medical imaging tests, one can ask if there are ways to shield parts of one's body from radiation that aren't being tested and making sure the doctor performing the test is qualified.[89]

Small doses of ionizing radiation increase risk by a very small amount. Most studies haven't detected increased cancer risk in people exposed to low levels of ionizing radiation. For example, people living in higher altitudes don't have noticeably higher cancer rates than other people. In general, cancer risk from radiation increases as the dose of radiation increases and there is thought to be no safe level of exposure. Ultraviolet radiation as a type of radiation that can be ionizing radiation. Sunlight is the main source of ultraviolet radiation.[84]

Factors that increase one's exposure to ultraviolet radiation when outside include:

- Time of day. Almost ⅓ of UV radiation hits the surface between 11AM and 1PM, and ¾ hit the surface between 9AM and 5PM.
- Time of year. UV radiation is greater during summer. This factor is less significant near the equator.
- Altitude. High elevation causes more UV radiation to penetrate the atmosphere.
- Clouds. Sometimes clouds decrease levels of UV radiation because they block UV radiation from the sun. Other times, they increase exposure because they reflect UV radiation.
- Reflection off surfaces, such as water, sand, snow, and grass increases UV radiation.
- Ozone density, because ozone stops some UV radiation from reaching the surface.

Some tips to decrease exposure to UV radiation:

- Stay in the shade. This is one of the best ways to limit exposure to UV radiation in sunlight.
- Cover yourself with clothing.
- Wear sunglasses.
- Use sunscreen on exposed skin.[90]

Tanning beds are also a source of ultraviolet radiation. Using tanning booths can increase one's chance of getting skin melanoma by at least 75%.[91]

Vitamin D_3 is also produced from ultraviolet radiation, although the American Society for Clinical Nutrition states that vitamin D is readily available from supplements and that the controversy about reducing ultraviolet radiation exposure was fueled by the tanning industry.[92]

There could be some risk of cell phone use being associated with cancer, but the evidence is not strong enough to be considered causal and needs to be investigated further.[93, 118]

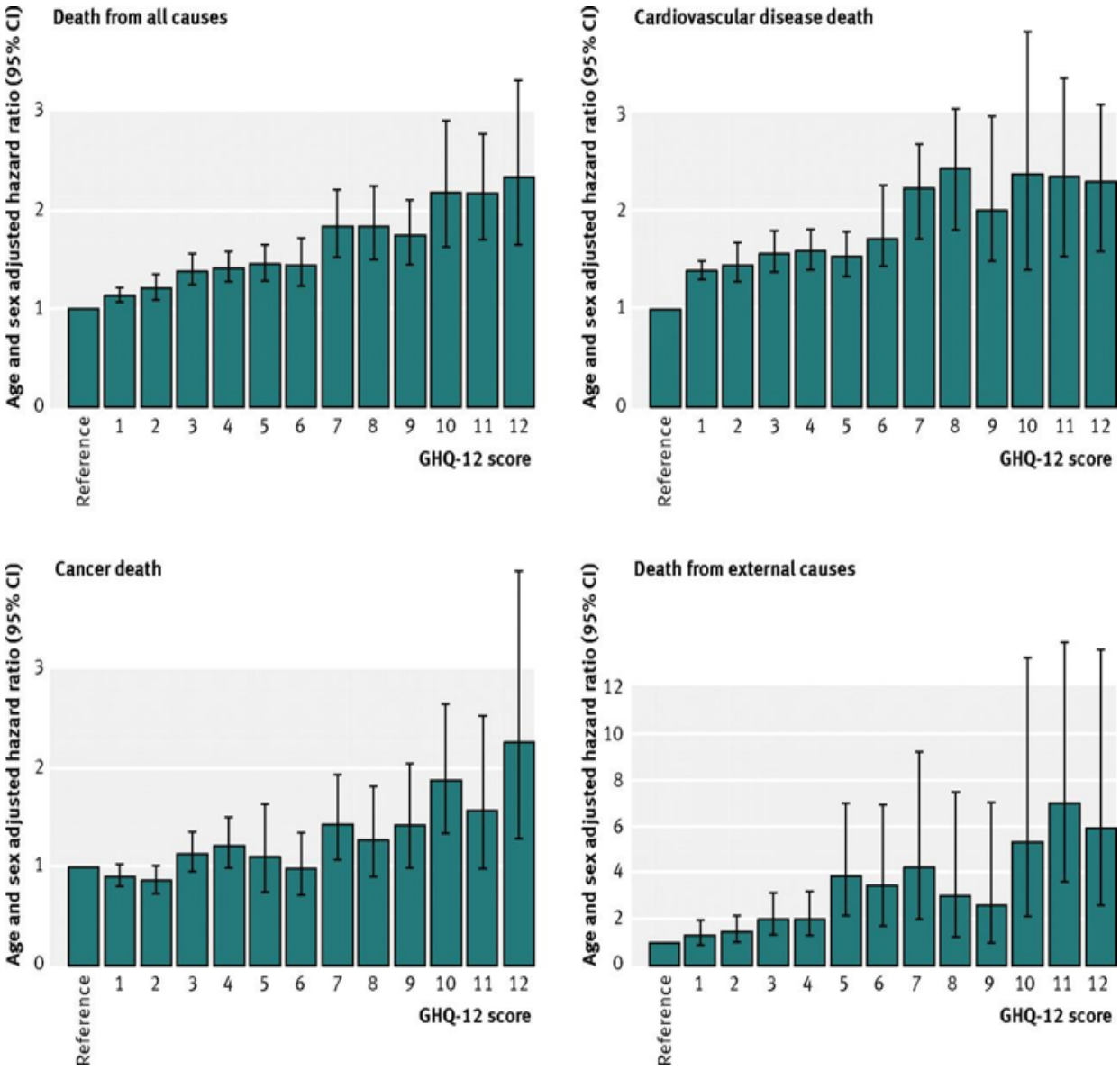
Emotions and feelings

Positive emotions and feelings

A review suggested that positive emotions and feelings decreased mortality. Proposed mechanisms include positive emotions and feelings being associated with better health practices such as improved sleep quality, increased exercise, and increased dietary zinc consumption, as well as lower levels of some stress hormones. It has also been hypothesized to be associated with other health-relevant hormones, various aspects of immune function, and closer and more social contacts.[33] Less Wrong has a good article on how to be happy.

Psychological distress

A meta-analysis was conducted on psychological stress. To measure psychological stress, it used the GHQ-12 score, which measured symptoms of anxiety, depression, social dysfunction, and loss of confidence. The scores range from 0 to 12, with 0 being asymptomatic, 1-3 being subclinically symptomatic, 4-6 being symptomatic, and 7-12 being highly symptomatic. It found the results shown in the following graphs.



This association was essentially unchanged after controlling for a range of covariates including occupational social class, alcohol intake, and smoking. However, reverse causality may still partly explain the association.[30]

Stress

A study found that individuals with moderate and high stress levels as opposed to low stress had hazard ratios (HRs) of mortality of 1.43 and 1.49, respectively.[27] A meta-analysis found that high perceived stress as opposed to low perceived stress had a coronary heart disease relative risk (RR) of 1.27. The mean age of participants in the studies used in the meta-analysis varied from 44 to 72.5 years and was significantly and positively associated with effect size. It explained 46% of the variance in effect sizes between the studies used in the meta-analysis.[28]

A cross-sectional study (which is a relatively weak study design) not in the aforementioned meta-analysis used 28,753 subjects to study the effect on mortality from the amount of stress and the perception of whether stress is harmful or not. It found that neither of these factors predicted mortality independently, but but that taken together, they did have a statistically significant effect. Subjects who reported much stress *and* that stress has a large effect on health had a HR of 1.43 (95% CI: 1.2, 1.7). Reverse causality may partially explain this though, as those who have had negative health impacts from stress may have been more likely to report that stress influences health.[83]

Anger and hostility

A meta-analysis found that after fully controlling for behavior covariates such as smoking, physical activity or body mass index, and socioeconomic status, anger and hostility was not associated with coronary heart disease (CHD), though the results are inconclusive.[34]

Social and personality factors

Social status

A review suggested that social status is linked to health via gender, race, ethnicity, education levels, socioeconomic differences, family background, and old age.[46]

Giving to others

An observational study found that stressful life events was not a predictor for mortality for those who engaged in unpaid helping behavior directed towards friends, neighbors, or relatives who did not live with them. This association may be due to giving to others causing one to have a sense of mattering, opportunities for generativity, improved social well-being, the emotional state of compassion, and the physiology of the caregiving behavioral system.[35]

Social relationships

A large meta-analysis found that the odds ratio of mortality of having weak social relationships is 1.5 (95% confidence interval (CI): 1.42 to 1.59). However, this effect may be a conservative estimate. Many of the studies used in the meta-analysis used single item measures of social relations, but the size of the association was greatest in studies that used more complex measurements. Additionally, some of the studies in the meta-analysis adjusted for risk factors that may be mediators of social relationships' effect on mortality (e.g. behavior, diet, and exercise). Many of the studies in the meta-analysis also ignored the quality of social relationships, but research suggests that negative social relationships are linked to increased mortality. Thus, the effect of social relationships on mortality could be even greater than the study found.

Concerning causation, social relationships are linked to better health practices and psychological processes, such as stress and depression, which influence health outcomes on their own. However, the meta-analysis also states that social relationships exert an independent effect. Some studies show that social support is linked to better immune system functioning and to immune-mediated inflammatory processes.[36]

Conscientiousness

A cohort study with 468 deaths found that each 1 standard deviation decrease in conscientiousness was associated with HR being multiplied by 1.07 (95% CI: 0.98 – 1.17), though it gave no mechanism for the association.[39] Although it adjusted for several variables, (e.g. socioeconomic status, smoking, and drinking), it didn't adjust for drug use, risky driving, risky sex, suicide, and violence, which were all found by a meta-analysis to have statistically significant associations with conscientiousness.[40] Overall, it seems to me that conscientiousness doesn't seem to have a significant effect on mortality.

Infectious diseases

Mayo clinic has a good article on preventing infectious disease.

Dental health

A cohort study of 5611 adults found that compared to men with 26-32 teeth, men with 16-25 teeth had an HR of 1.03 (95% CI: 0.91-1.17), men with 1-15 teeth had an HR of 1.21 (95% CI: 1.05-1.40) and men with 0 teeth had an HR of 1.18 (95% CI: 1.00-1.39).

In the study, men who never brushed their teeth at night had a HR of 1.34 (95% CI: 1.14-1.57) relative to those who did every night. Among subjects who brushed at night, HR was similar between those who did and didn't brush daily in the morning or day. The HR for men who brushed in the morning every day but not at night every day was 1.19 (95% CI: 0.99-1.43).

In the study, men who never used dental floss had an HR of 1.27 (95% CI: 1.11-1.46) and those who sometimes used it had an HR or 1.14 (95% CI: 1.00-1.30) compared to men who used it every day. Among subjects who brushed their teeth at night daily, not flossing was associated with a significantly increased HR.

Use of toothpicks didn't significantly decrease HR and mouthwash had no effect.

The study had a list of other studies on the effect of dental health on mortality. It seems to us that almost all of them found a negative correlation between dental health and risk of mortality, although the study didn't say their methodology for selecting the studies to show. I did a crude review of other literature by only looking at their abstracts and found that five studies found that poor dental health increased risk of mortality and one found it didn't.

Regarding possible mechanisms, the study says that toothpaste helps prevent dental caries and that dental floss is the most effective means of removing interdental plaque and decreasing interdental gingival inflammation.[38]

Sleep

It seems that getting too little or too much sleep likely increases one's risk of mortality, but it's hard to tell exactly how much is too much and how little is too little.

One review found that the association between amount of sleep and mortality is inconsistent in studies and that what association does exist may be due to reverse-causality.[41] However, a meta-analysis found that the RR associated with short sleep duration (variously defined as sleeping from < 8 hrs/night to < 6 hrs/night) was 1.10 (95% CI: 1.06-1.15). It also found that the RR associated with long sleep duration (variously defined as sleeping for > 8 hrs/night to > 10 hrs per night) compared with medium sleep duration (variously defined as sleeping for 7-7.9 hrs/night to 9-9.9 hrs/night) was 1.23 (95% CI: 1.17 - 1.30).[42]

The National Heart, Lung, and Blood Institute and Mayo Clinic recommend adults get 7-8 hours of sleep per night, although it also says sleep needs vary from person to person. It gives no method of determining optimal sleep for an individual. Additionally, it doesn't say if its recommendations are for optimal longevity, optimal productivity, something else, or a combination of factors.[43] The Harvard Medical School implies that one's optimal amount of sleep is enough sleep to not need an alarm to wake up, though it didn't specify the criteria for determining optimality either.[45]

Drugs

None of the drugs I've looked into have a beneficial effect for the people without a special disease or risk factor. Notes on them are here.

Blood donation

A quasi-randomized experiment with a validity near that of a randomized trial presumably suggested that blood donation didn't significantly decrease risk of coronary heart disease (CHD). Observational studies have shown much lower CHD incidence among donors, although the authors of the former experiment suspect that bias and reverse causation played a role in this.[29] That said, a review found that reverse causation accounted for only 30% of the effect of blood donation, though I haven't been able to find the review. RomeoStevens suggests that the potential benefits of blood donation are high enough and the costs are low enough that blood donation is worth doing.[120]

Sitting

After adjusting for amount of physical activity, a meta-analysis estimated that for every one hour increment of sitting in intervals 0-3, >3-7 and >7 h/day total sitting time, the hazard ratios of mortality were 1.00 (95% CI: 0.98-1.03), 1.02 (95% CI: 0.99-1.05) and 1.05 (95% CI: 1.02-1.08) respectively. It proposed no mechanism for sitting time having this effect,[37] so it might have been due to confounding variables it didn't control.

Sleep apnea

Sleep apnea is an independent risk factor for mortality and cardiovascular disease.[26] Symptoms and other information on sleep apnea are here.

Snoring

A meta-analysis found that self-reported habitual snoring had a small but statistically significant association with stroke and coronary heart disease, but not with cardiovascular disease and all-cause mortality [HR 0.98 (95% CI: 0.78-1.23)]. Whether the risk is due to obstructive sleep apnea is controversial. Only the abstract is able to be viewed for free, so I'm just basing this off the abstract.[31]

Exams

The organization Susan G. Komen, citing a meta-analysis that used randomized controlled trials, doesn't recommend breast self exams as a screening tool for breast cancer, as it hasn't been shown to decrease cancer death. However, it still stated that it is important to be familiar with one's breasts' appearance and how they normally feel.[49] According to the Memorial Sloan Kettering Cancer Center, no study has been able to show a statistically significant decrease in breast cancer deaths from breast self-exams.[50] The National Cancer Institute states that breast self-examinations haven't been shown to decrease breast cancer mortality, but does increase biopsies of benign breast lesions.[51]

The American Cancer Society doesn't recommend testicular self-exams for all men, as they haven't been studied enough to determine if they decrease mortality. However, it states that men with risk factors of testicular cancer (e.g. an undescended testical, previous testicular cancer, of a family member who previously had testicular cancer) should consider self-exams and discuss them with a doctor. The American Cancer Society also recommends having testicular self-exams in routine cancer-related check-ups.[52]

Genomics

Genomics is the study of genes in one's genome, and may help increase health by using knowledge of one's genes to have personalized treatment. However, it hasn't proved to be useful for most; recommendations rarely change after knowledge from genomic testing. Still, genomics has much future potential.[102]

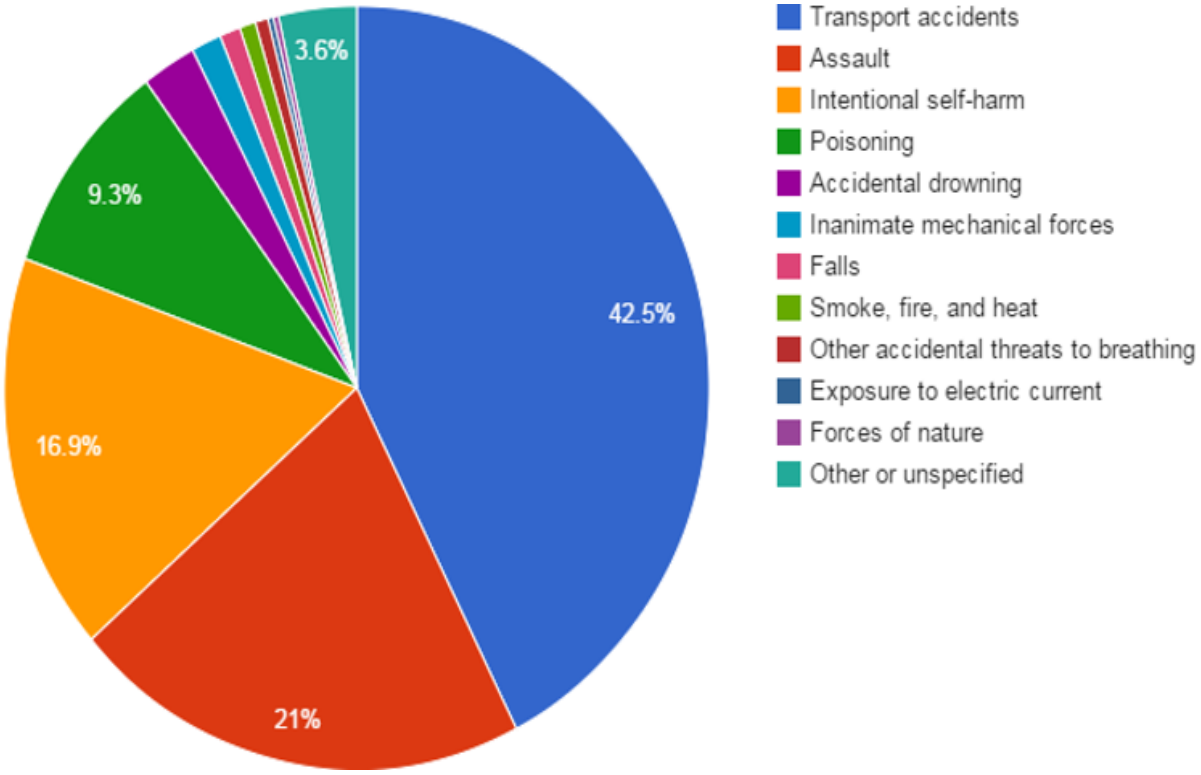
Aging

Like I've said in the section "Can we become immortal," the proportion of deaths that are caused by aging in the industrial world approaches *90%*,[53] but some organizations and companies are working on curing it.[54, 55, 56]

One could support these organizations in an effort to hasten the development of anti-aging therapies, although I doubt an individual would have a noticeable impact on one's own chance of death unless one is very wealthy. That said, I have little knowledge in investments, but I suppose investing in companies working on curing aging may be beneficial, as if they succeed, they may offer an enormous return on investment, and if they fail, one would probably die, so losing one's money may not be as bad. Calico currently isn't a public stock, though.

External causes of death

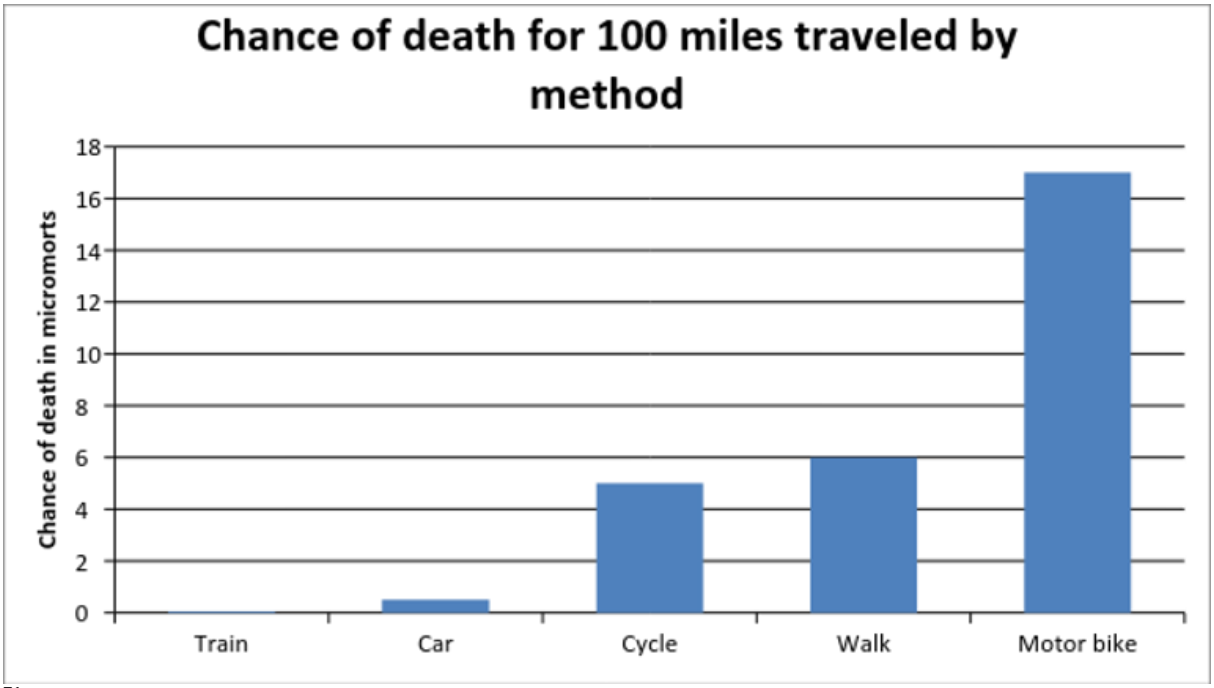
Unless otherwise specified, graphs in this section are on data collected from American citizens ages 15-24, as based off the Less Wrong census results, this seems to be the most probable demographic that will read this. For this demographic, external causes cause 76% of deaths. Note that although this is true, one is much more likely to die when older than when aged 15-24, and older individuals are much more likely to die from disease than from external causes of death. Thus, I think it's more important when young to decrease risk of disease than external causes of death. The graph below shows the percentage of total deaths from external causes caused by various causes.



21

Transport accidents

Below are the relative death rates of specified means of transportation for people in general:



71

Much information about preventing death from car crashes is here. Information on preventing death from car crashes is here, here, here, and here.

Assault

Lifehacker's "Basic Self-Defense Moves Anyone Can Do (and Everyone Should Know)" gives a basic introduction to self defence.

Intentional self harm

Intentional self harm such as suicide, presumably, increases one's risk of death.[47] Mayo Clinic has a guide on preventing suicide. I recommend looking at it if you are considering killing yourself. Additionally, if are are considering killing yourself, I suggest reviewing the potential rewards of achieving immortality from the section "Should we try to become immortal."

Poisoning

What to do if a poisoning occurs

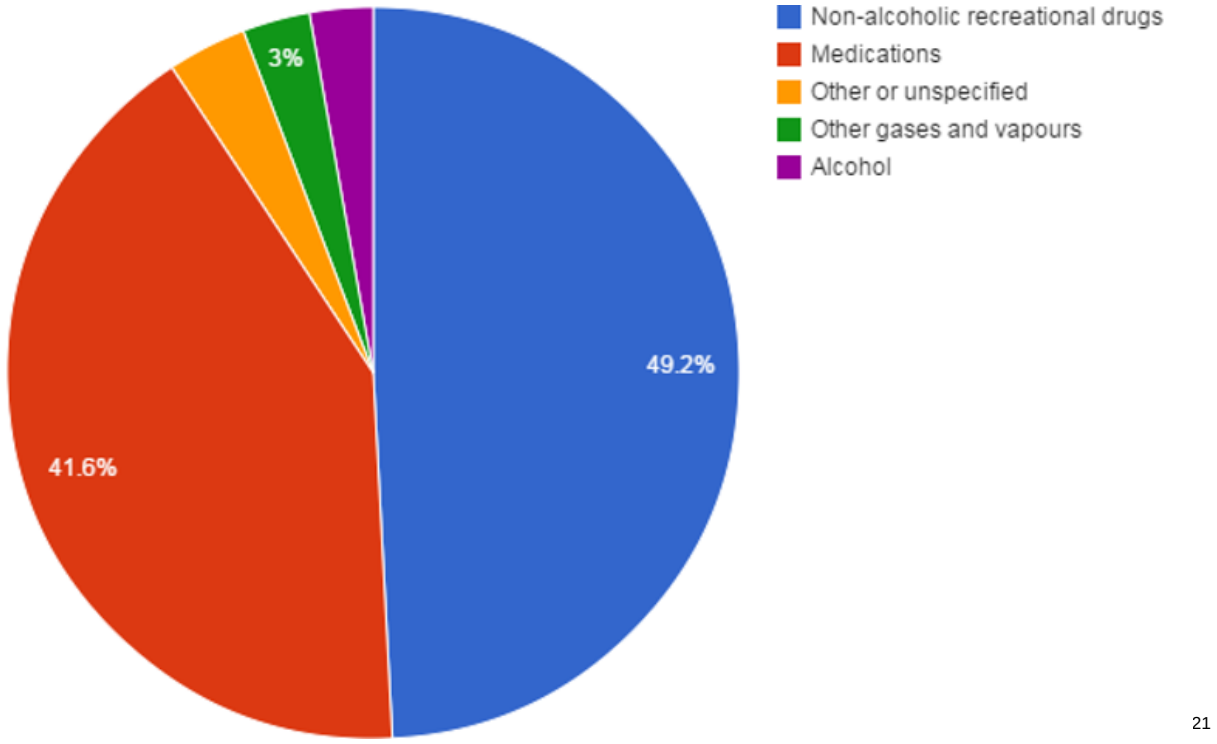
CDC recommends staying calm, dialing 1-800-222-1222, and having this information ready:

- Your age and weight.
- If available, the container of the poison.
- The time of the poison exposure.
- The address where the poisoning occurred.

It also recommends staying on the phone and following the instructions of the emergency operator or poison control center.[18]

Types of poisons

Below is a graph of the risk of death per type of poison.



21

Some types of poisons:

- Medicine overdoses.
- Some household chemicals.
- Recreational drug overdoses.
- Carbon monoxide.
- Metals such as lead and mercury.
- Plants[12] and mushrooms.[14]
- Presumably some animals.
- Some fumes, gases, and vapors.[15]

Recreational drugs

Using recreational drugs increases risk of death.

Medicine overdoses and household chemicals

CDC has tips for these here.

Carbon monoxide

CDC and Mayo Clinic have tips for this here and here.

Lead

Lead poisoning causes 0.2% of deaths worldwide and *0.0%* of deaths in developed countries.[22] Children under the age of 6 are at higher risk of lead poisoning.[24] Thus, for those who aren't children, learning more about preventing lead poisoning seems like more effort than it's worth. No completely safe blood lead level has been identified.[23]

Mercury

MedlinePlus has an article on mercury poisoning here.

Accidental drowning

Information on preventing accidental drowning from CDC is here and here.

Inanimate mechanical forces

Over half of deaths from inanimate mechanical forces for Americans aged 15-24 are from firearms. Many of the other deaths are from explosions, machinery, and getting hit by objects. I suppose using common sense, precaution, and standard safety procedures when dealing with such things is one's best defense.

Falls

Again, I suppose common sense and precaution is one's best defense. Additionally, alcohol and substance abuse is a risk factor of falling.[72]

Smoke, fire and heat

Owning smoke alarms halves one's risk of dying in a home fire.[73] Again, common sense when dealing with fires and items potentially causing fires (e.g. electrical wires and devices) seems effective.

Other accidental threats to breathing

Deaths from other accidental threats to breathing are largely caused by strangling or choking on food or gastric contents, and occasionally by being in a cave-in or trapped in a low-oxygen environment.[21] Choking can be caused by eating quickly or laughing while eating.[74] If you are choking:

- Forcefully cough. Lean as far forwards as you can and hold onto something that is firmly anchored, if possible. Breathe out and then take a deep breath in and cough; this may eject the foreign object.
- Attract someone's attention for help.[75]

Additionally, choking can be caused by vomiting while unconscious, which can be caused by being very drunk.[76] I suggest lying in the recovery position if you think you may vomit while unconscious, so as to to decrease the chance of choking on vomit.[77] Don't forget to use common sense.

Electric current

Electric shock is usually caused by contact with poorly insulated wires or ungrounded electrical equipment, using electrical devices while in water, or lightning.[78] Roughly ⅓ of deaths from electricity are caused by exposure to electric transmission lines.[21]

Forces of nature

Deaths from forces of nature in (for Americans ages 15-24) in descending order of number of deaths caused are: exposure to cold, exposure to heat, lightning, avalanches or other earth movements, cataclysmic storms, and floods.[21] Here are some tips to prevent these deaths:

- When traveling in cold weather, carry emergency supplies in your car and tell someone where you're heading.[79]
- Stay hydrated during hot weather.[80]
- Safe locations from lightning include substantial buildings and hard-topped vehicles. Safe locations don't include small sheds, rain shelters, and open vehicles.
- Wait until there are no thunderstorm clouds in the area before going to a location that isn't lightning safe.[81]

Medical care

Since medical care is tasked with treating diseases, receiving medical care when one has illnesses presumably decreases risk of death. Though necessary medical care may be essential when one has illnesses, a review estimated that preventable medical errors contributed to roughly 440,000 deaths per year in the US, which is *roughly one-sixth of total deaths in the US*. It gave a lower limit of 210,000 deaths per year.

The frequency of deaths from preventable medical errors varied across studies used in the review, with a hospital that was shown the put much effort into improving patient safety having a lower proportion of deaths from preventable medical errors than that of others.[57] Thus, I suppose that it would be beneficial to go to hospitals that are known for their dedication to patient safety. There are several rankings of hospital safety available on the internet, such as this one. Information on how to help prevent medical errors is found here and under the "What Consumers Can Do" section here. One rare medical error is having a surgery be done on the wrong body part. The New York Times gives tips for preventing this here.

Additionally, I suppose it may be good to live relatively close to a hospital so as to be able to quickly reach it in emergencies, though I've found no sources stating this.

A common form of medical care are general health checks. A comprehensive Cochrane review with 182,880 subjects concluded that general health checks are probably not beneficial.[107] A meta-analysis found that general health checks are associated with small but statistically significant benefits in factoring related to mortality, such as blood pressure and body mass index. However, it found no significant association with mortality.[109] The New York Times acknowledged that health checks are probably not beneficial and gave some explanation why general health checks are nonetheless still common.[108] However, CDC and MedlinePlus recommend getting routine general health checks. The cited no studies to support their claims.[104, 106] When I contacted CDC about it, it responded, "Regular health exams and tests can help find problems before they start. They also can help find problems early, when your chances for treatment and cure are better. By getting the right health services, screenings, and treatments, you are taking steps that help your chances for living a longer, healthier life," a claim that doesn't seem supported by evidence. It also stated, "Although CDC understands you are concerned, the agency does not comment on information from unofficial or non-CDC sources." I never heard back from MedlinePlus.

Cryonics

Cryonics is the freezing of legally dead humans with the purpose preserving their bodies so they can be brought back to life in the future once technology makes it possible. Human tissue have been cryopreserved and then brought back to life, although this has never been done on full humans.[59] The price of Cryonics at least ranges from $28,000 to $200,000.[60] More information on cryonics is on LessWrong Wiki.

Money

Cryonics, medical care, safe housing, and basic needs all take money. Rejuvenation therapy may also be very expensive. It seems valuable to have a reasonable amount of money and income.

Future advancements

Keeping updated on further advancements in technology seems like a good idea, as not doing so would prevent one from making use of future technologies. Keeping updated on advancements on curing aging seems especially important, due to the massive

number of casualties it inflicts and the current work being done to stop it. Updates on mind-uploading seem important as well. I don't know of any very efficient method of keeping updated on new advancements, but periodically googling for articles about curing aging or Calico and searching for new scientific articles on topics in this guide seems reasonable. As knb suggested, it seems beneficial to periodically check on *Fight Aging*, a website advocating anti-aging therapies. I'll try to do this and update this guide with any new relevant information I find.

There is much uncertainty ahead, but if we're clever enough, we just might make it though alive.

References

1. Actual Causes of Death in the United States, 2000.
2. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care.
3. All pages in The Nutrition Source, a part of the Harvard School of Public Health.
4. Will calorie restriction work on humans?
5. The pages Getting Started, Tests and Biomarkers, and Risks from The CR Society.
6. The causal role of breakfast in energy balance and health: a randomized controlled trial in lean adults.
7. Low Glycemic Index: Lente Carbohydrates and Physiological Effects of altered food frequency. Published in 1994.
8. Leisure Time Physical Activity of Moderate to Vigorous Intensity and Mortality: A Large Pooled Cohort Analysis.
9. Exercising for Health and Longevity vs Peak Performance: Different Regimens for Different Goals.
10. Water: How much should you drink every day?
11. MET-hour equivalents of various physical activities.
12. Poisoning. NLM
13. Carcinogen. Dictionary.com
14. Types of Poisons. New York Poison Center
15. The Most Common Poisons for Children and Adults. National Capital Poison Center.
16. Known and Probable Human Carcinogens. American cancer society.
17. Nutritional Effects of Food Processing. Nutritiondata.com.
18. Tips to Prevent Poisonings. CDC.
19. Carbon monoxide poisoning. Mayo Clinic.
20. Carbon Monoxide Poisoning. CDC.
21. CDCWONDER. Query Criteria taken from all genders, all states, all races, all levels of urbanization, all weekdays, dates 1999 – 2010, ages 15 – 24.
22. Global health risks: mortality and burden of disease attributable to selected major risks.
23. National Biomonitoring Program Factsheet. CDC
24. Lead poisoning. Mayo Clinic.
25. Mercury. Medline Plus.
26. Snoring Is Not Associated With All-Cause Mortality, Incident Cardiovascular Disease, or Stroke in the Busselton Health Study.
27. Do Stress Trajectories Predict Mortality in Older Men? Longitudinal Findings from the VA Normative Aging Study.
28. Meta-analysis of Perceived Stress and its Association with Incident Coronary Heart Disease.
29. Iron and cardiac ischemia: a natural, quasi-random experiment comparing eligible with disqualified blood donors.
30. Association between psychological distress and mortality: individual participant pooled analysis of 10 prospective cohort studies.

31. Self-reported habitual snoring and risk of cardiovascular disease and all-cause mortality.
32. Is it true that occasionally following a fasting diet can reduce my risk of heart disease?
33. Positive Affect and Health.
34. The Association of Anger and Hostility with Future Coronary Heart Disease: A Meta-Analytic Review of Prospective Evidence.
35. Giving to Others and the Association Between Stress and Mortality.
36. Social Relationships and Mortality Risk: A Meta-analytic Review.
37. Daily Sitting Time and All-Cause Mortality: A Meta-Analysis.
38. Dental Health Behaviors, Dentition, and Mortality in the Elderly: The Leisure World Cohort Study.
39. Low Conscientiousness and Risk of All-Cause, Cardiovascular and Cancer Mortality over 17 Years: Whitehall II Cohort Study.
40. Conscientiousness and Health-Related Behaviors: A Meta-Analysis of the Leading Behavioral Contributors to Mortality.
41. Sleep duration and all-cause mortality: a critical review of measurement and associations.
42. Sleep duration and mortality: a systematic review and meta-analysis.
43. How Much Sleep Is Enough? National Lung, Blood, and Heart Institute.
44. How many hours of sleep are enough for good health? Mayo Clinic.
45. Assess Your Sleep Needs. Harvard Medical School.
46. A Life-Span Developmental Perspective on Social Status and Health.
47. Suicide. Merriam-Webster.
48. Can testosterone therapy promote youth and vitality? Mayo Clinic.
49. Breast Self-Exam. Susan G. Komen.
50. Screening Guidelines. The Memorial Sloan Kettering Cancer Center.
51. Breast Cancer Screening Overview. The National Cancer Institute.
52. Testicular self-exam. The American Cancer Society.
53. Life Span Extension Research and Public Debate: Societal Considerations.
54. SENS Research Foundation: About.
55. Science for Life Extension Homepage.
56. Google's project to 'cure death,' Calico, announces $1.5 billion research center. The Verge.
57. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care.
58. When Surgeons Cut the Wrong Body Part. The New York Times.
59. Cold facts about cryonics. The Guardian.
60. The cryonics organization founded by the "Father of Cryonics," Robert C.W. Ettinger. Cryonics Institute.
61. Escape Velocity: Why the Prospect of Extreme Human Life Extension Matters Now.
62. International Journal of Machine Consciousness Introduction.
63. The Philosophy of 'Her.' The New York Times.
64. How to Survive the End of the Universe. Discover Magazine.
65. A Space-Time Crystal to Outlive the Universe. Universe Today.
66. Conjunction Fallacy. Less Wrong.
67. Cognitive Biases Potentially Affecting Judgment of Global Risks.
68. Genetic influence on human lifespan and longevity.
69. First Drug Shown to Extend Life Span in Mammals. MIT Technology Review.
70. Sirolimus (Oral Route). Mayo Clinic.
71. Micromorts. Understanding Uncertainty.
72. Falls. WHO.
73. Smoke alarm outreach materials. US Fire Administration.
74. What causes choking? 17 possible conditions. Healthline.
75. Choking. Better Health Channel.
76. Aspiration pneumonia. HealthCentral.
77. First aid - Recovery position. NHS Choices.
78. Electric Shock. HowStuffWorks.

79. Hypothermia prevention. Mayo Clinic.
80. Extreme Heat: A Prevention Guide to Promote Your Personal Health and Safety. CDC.
81. Understanding the Lightning Threat: Minimizing Your Risk. National weather service.
82. The Case Against QuikClot. The survival mom.
83. Does the Perception that Stress Affects Health Matter? The Association with Health and Mortality.
84. Cancer Prevention. WHO.
85. Infections That Can Lead to Cancer. American Cancer Society.
86. Pollution. American Cancer Society.
87. Occupations or Occupational Groups Associated with Carcinogen Exposures. Canadian Centre for Occupational Health and Safety.
88. Radon. American Cancer Society.
89. Medical radiation. American Cancer Society.
90. Ultraviolet (UV) Radiation. American Cancer Society.
91. An Unhealthy Glow. American Cancer Society.
92. Sun exposure and vitamin D sufficiency.
93. Cell Phones and Cancer Risk. National Cancer Institute.
94. Nutrition for Everyone. CDC.
95. How Can I Tell If My Body is Missing Key Nutrients? Oprah.com.
96. Decaffeination, Green Tea and Benefits. Teas etc.
97. Red and Processed Meat Consumption and Risk of Incident Coronary Heart Disease, Stroke, and Diabetes Mellitus.
98. Lifestyle interventions to increase longevity.
99. Chemicals in Meat Cooked at High Temperatures and Cancer Risk. National Cancer Institute.
100. Are You Living in a Simulation?
101. How reliable are scientific studies?
102. Genomics: What You Should Know. Forbes.
103. Organic foods: Are they safer? More nutritious? Mayo Clinic.
104. Health screening - men - ages 18 to 39. MedlinePlus.
105. Why do I need medical checkups. Banner Health.
106. Regular Check-Ups are Important. CDC.
107. General health checks in adults for reducing morbidity and mortality for disease (Review).
108. Let's (Not) Get Physicals.
109. Effectiveness of general practice-based health checks: a systematic review and meta-analysis.
110. Supplements: Nutrition in a Pill? Mayo Clinic.
111. Nutritional Effects of Food Processing. SelfNutritionData.
112. What Is the Healthiest Drink? SFGate.
113. Leading Causes of Death. CDC.
114. Bias Detection in Meta-analysis. Statistical Help.
115. The summary of Sodium Intake in Populations: Assessment of Evidence. Institute of Medicine.
116. Compared With Usual Sodium Intake, Low and Excessive -Sodium Diets Are Associated With Increased Mortality: A Meta-analysis.
117. The Cochrane Review of Sodium and Health.
118. Is there any link between cellphones and cancer? Mayo Clinic.
119. A glass of red wine a day keeps the doctor away. Yale-New Haven Hospital.
120. Comment on Lifestyle Interventions to Increase Longevity. Less Wrong.

An Introduction to Control Theory

Behavior: The Control of Perception by William Powers applies control theory to psychology to develop a model of human intelligence that seems relevant to two of LW's primary interests: effective living for humans and value-preserving designs for artificial intelligence. It's been discussed on LW previously here, here, and here, as well as mentioned in Yvain's roundup of 5 years (and a week) of LW. I've found previous discussions unpersuasive for two reasons: first, they typically only have a short introduction to control theory and the mechanics of control systems, making it not quite obvious what specific modeling techniques they have in mind, and second, they often fail to communicate the differences between this model and competing models of intelligence. Even if you're not interested in its application to psychology, control theory is a widely applicable mathematical toolkit whose basics are simple and well worth knowing.

Because of the length of the material, I'll split it into three posts. In this post, I'll first give an introduction to that subject that's hopefully broadly accessible. The next post will explain the model Powers introduces in his book. In the last post, I'll provide commentary on the model and what I see as its implications, for both LW and AI.

Control theory is a central tool of modern engineering. Briefly, most interesting things can be modeled as dynamical systems, having both states and rules on how those states change with time. Consider the 3D position and velocity of a ball in a bowl (with friction); six numbers tell you where the ball is, its speed, and its direction of movement, and a formula tells you how you can predict what those six numbers will be in the next instant given where they are now. Those systems can be characterized by their **attractors**, states that are stable and are the endpoint for nearby states. The ball sitting motionless in the bottom of the bowl is an attractor- if it's already there, it will stay there, and if it's nearby (releasing from a centimeter away, for example), it will eventually end up there. The point of control theory is that adding a **control** to a dynamical system allows you to edit the total system dynamics so that the points you *want* to be stable attractors *are* stable attractors.

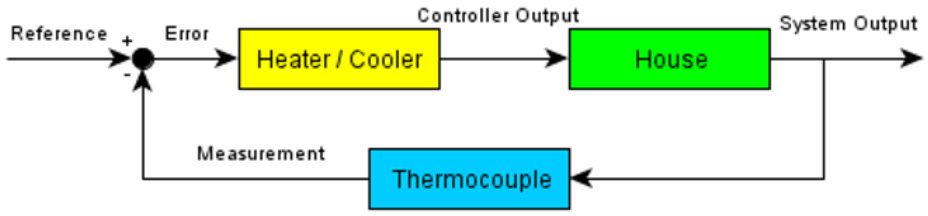
Let's flesh out that sketch with an example. Suppose you want to keep a house within a specific temperature range. You have a sensor of the current temperature, a heater, and a cooler. A thermostat takes the sensor's output, compares it to the desired temperature range, and turns the heater on if the sensed temperature is below the desired temperature range, and turns if off if the sensed temperature is above the minimum of that range, and does the reverse with the cooler (turning it on if the sensor is above the desired max, and turning it off if it's below).

Most interesting control systems are have a finer range of control values- instead of simply flipping a switch on or off, a car's cruise control can smoothly vary the amount of gas or brake that's being applied. A simple way to make a control system is to take the difference between the desired speed and actual speed, multiply it by some factor to go from units of distance per time to angle of pedal, and adjust the position of the pedals accordingly. (If the function is linear, it's called a linear controller.)

Let's introduce more of the technical vocabulary. The thing we're measuring is the **input** (to the controller), the level we want it to be at is the **reference**, the difference between those is the **error**, and the adjustment the control system makes is the

output or **feedback** (sometimes we'll talk about the **actuator** as the physical means by which the controller emits its output). None of them have to be single variables- they can be vectors, which allow us to describe arbitrarily complicated systems. (The six numbers that express the position and velocity of a ball, each in three dimensions, are an example of an input vector.) I'll often use the noun **state** to refer to the, well, *state* of the system, and 'points in state-space' refers to those states as vectors. There's also a possible confusion in that the *plant* (the system being controlled) and the *controller* are mirrored- the controller's output is the plant's input, and the plant's output is the controller's input.

Control systems naturally lend themselves to diagrams: here's the block diagram from the thermostat and cruise control:



In a block diagram, each block is some function of its inputs, and the arrows show what affects what. Moving left to right, the reference is the temperature you've set, the current temperature is subtracted (that's the point of the plus and minus sign), and the error goes into the yellow box which represents the function that converts from the error to the effort put into altering the system. That's the controller output arrow that goes into the green box (the house), which represents the external system. This is also a functional block, because it takes the controller output and any disturbances (often represented as another arrow pointing in from the top) and converts them into the system temperature. The arrow leading out of the house points both to the right- to remind you that this is the temperature you're living with- and back into the thermocouple, the sensor that measures the temperature to compare with the reference, and now we've finished our feedback loop.

Now that we have a mathematical language for modeling lots of different systems, we can abstract away the specifics and prove properties about how those systems will behave given various controllers (i.e. feedback functions). Feedback functions convert the input and reference to the output, and are the mathematical abstraction of a physical controller. They can be arbitrary functions, but most of the mathematical edifice of control theory assumes that everything is continuous (but not necessarily linear). If you know the dynamics of a system, you can optimize your control function to match the system and be guaranteed to converge to the reference with a particular time profile. Rather than go deeply into the math, I'll discuss a few concepts that have technical meanings in control theory that are useful to think about.

First is **convergence**: the system output will eventually match the reference. This means that any errors that get introduced into the system are transient (temporary), and ideally we know the time profile of how large those errors will be as time progresses. A common goal here is **exponential** convergence, which means that the

error decreases with a rate proportional to its size. (If the temperature is off by 2 degrees, the rate at which the error is decreasing is twice that of the rate when the temperature is off by 1 degree.) A simple linear controller will, for simple state dynamics, accomplish exponential convergence. If your system doesn't converge, then you are not successfully controlling it, and if your reference changes unpredictably at a rate faster than your system can converge, then you are not going to be able to match your reference closely.

Second is **equilibrium**: a point when the forces are balanced. If the system is at an equilibrium state, then nothing will change. This is typically discussed along with steady state error: imagine that my house gets heated by the sun at a rate of 1° an hour, and rather than a standard air conditioner that's on or off I have a 'dimmer switch' for my AC. If my controller sets the AC rate at the difference between the reference temperature and the current temperature per hour, then when the house is at 30° and I want it to be at 23° it'll try to reduce the temperature by 7° an hour, but when the house is at 24° and I want it to be at 23° it will try to reduce the temperature by 1° an hour, which cancels the effect of the sun, and so the house is at equilibrium at 24°. (Most real controllers have an integrator in order to counteract this effect.)

Third is **stability**: even when we know a point is an equilibrium, we want to know about the behavior in the neighborhood of that point. A stable equilibrium is one where a disturbance will be corrected; an unstable equilibrium is one where a disturbance will be amplified. Imagine a pendulum with the bob balanced at the bottom- tap it and it'll eventually be balanced at the bottom again, because that's a stable equilibrium (also called an attractor). Now imagine the bob balanced at the top- tap it, and it'll race away from that point, unlikely to return, because that's an unstable equilibrium (also called a repulsor).

Stability has a second meaning in control theory: a controller that applies *too much* feedback will cause the system to go from a smaller positive error to a moderate negative error, and then again too much feedback will be applied, and the system will go from a moderate negative error to a huge positive error. Imagine a shower with a delay: you turn the temperature knob and five seconds later the temperature of the water coming out of the showerhead changes. If you react too strongly or too quickly, then you'll overreact, and the water that started out too hot will correct to being too cold by the time you've stopped turning the knob away from heat. That an overexuberant negative feedback controller can still lead to explosions is one of the interesting results of control theory, as is that making small, gradual, proportionate changes to the current state can be as effective as memory / implementing a delay in your controller. Sometimes, you achieve better control exactly because you applied *less* control effort.

It's also worth mentioning here that basically all real controllers (even amplifiers!) are negative feedback controllers- positive feedback leads to explosions (literally), because "positive feedback" in this context means "pushing the system in the direction of the error" rather than its meaning in psychology.

So what's the point of control systems? If we have a way of push the states of a system, we can effectively adjust the system dynamics to make *any* state that we want a stable state. If we put a motor at the joint of a pendulum, we can adjust the acceleration of the bob so that it has an equilibrium at one radian away from vertical, rather than zero radians away from vertical. If we put a thermostat in a house, we can

adjust the temperature changes of the house so that it returns to a comfortable range, instead of whatever the temperature is outside. (The point of control theory is to understand how the systems work, so we can make sure that our control systems do what we want them to do!)

Why are control systems interesting? Three primary reasons:

1. They're practical. Control systems are all over the place, from thermostats to cars to planes to satellites.
2. They can be adaptive. A **hierarchical** control system has a control loop which determines the parameters used in another control loop, and a natural application is adaptive control systems. When you launch a satellite, you might not have precise measurements of its rotational inertia, or you might expect that to change as it uses its fuel over its lifetime. One control system could observe how the satellite moves in response to its thrusters, and adjust the parameters of a rotational inertia model to correct errors to match the model to the observations. A second control system could use the inertia model created by the first control system to determine how to use the thrusters to adjust the satellite's alignment to match the desired rotation.
3. They give concrete mathematical models of how simple signal processing can create 'intentional' behavior, and of what it looks like to be intentional without an explicit model of reality. A centrifugal governor is not an agent in the LW sense of the term, but it is an agent in another sense of the term--an entity that performs actions on behalf of another. The governor is just some pieces of metal, it doesn't have a mind, it's not viewed with moral concern, it doesn't have goals as humans think of them, and it doesn't have even a rudimentary internal model of the dynamical system it's controlling, and it *still* gets the job done. Controllers seem like a class of entities that are best modeled by intentionality, in that they alter the state of their external environment to match their desired internal environment based on their perceptions, and can do so in arbitrarily complicated and powerful ways, but while they do steer the future they don't seem to be cross-domain and they don't look like anthropomorphic models of "human intelligence."

Next, B:CP on the use of control theory in psychology.

'Dumb' AI observes and manipulates controllers

The argument that AIs provided with a reward channel will observe their controllers and learn to manipulate them is a valid one. Unfortunately, it's often framed in a way that feels counterintuitive or extreme, especially to AI designers. It typically starts with the standard reinforcement learning scenario, then posits that the AI becomes superintelligent and either manipulates the controller with super-social powers, or breaks out and gains control of its reward channel, killing or threatening its controllers.

And that is a fair argument. But conceptually, it leaps from a standard reinforcement learning scenario, to a science-fiction-sounding scenario. It might help to have intermediate scenarios: to show that even lower intelligence AIs might start exhibiting the same sort of behaviour, long before it gets to superintelligence.

So consider the following scenario. Some complex, trainable AI is tasked with writing automated news stories for a student newspaper. It trawls the web and composes its stories, then gets reward and feedback from the editors. Assume there are two editors for this newspaper, and they work on alternate days. The two editors have somewhat different ideas as to what constitutes a good story, so their feedbacks are different. After a while, the AI finds that it gets higher reward by using a certain style on Monday, Wednesday and Friday, and another style on Tuesday and Thursdays - this is a simple consequence of its reward mechanism.

But the rota isn't perfect. Sometime the Monday editor will edit a story so late on Monday that it's Tuesday, and sometimes the Tuesday editor will be up early and edit a story at the same time. This pattern is more common during the holidays, when the editors go to different time zones, often travel, and become more irregular in their editing.

So if the AI complex and skilled enough, then, simply through simple feedback, it will start building up a picture of its editors. It will figure out when they are likely to stick to a schedule, and when they will be more irregular. It will figure out the difference between holidays and non-holidays. Given time, it may be able to track the editors moods and it will certainly pick up on any major change in their lives - such as romantic relationships and breakups, which will radically change whether and how it should present stories with a romantic focus.

It will also likely learn the correlation between stories and feedbacks - maybe presenting a story define roughly as "positive" will increase subsequent reward for the rest of the day, on all stories. Or maybe this will only work on a certain editor, or only early in the term. Or only before lunch.

Thus the simple trainable AI with a particular focus - write automated news stories - will be trained, through feedback, to learn about its editors/controllers, to distinguish them, to get to know them, and, in effect, to manipulate them.

This may be a useful "bridging example" between standard RL agents and the superintelligent machines.

Overpaying for happiness?

Happy New Year, everyone!

In the past few months I've been thinking several thoughts that all seem to point in the same direction:

1) People who live in developed Western countries usually make and spend much more money than people in poorer countries, but aren't that much happier. It feels like we're overpaying for happiness, spending too much money to get a single bit of enjoyment.

2) When you get enjoyment from something, the association between "that thing" and "pleasure" in your mind gets stronger, but at the same time it becomes less sensitive and requires more stimulus. For example if you like sweet food, you can get into a cycle of eating more and more food that's sweeter and sweeter. But the guy next door, who's eating much less and periodically fasting to keep the association fresh, is actually getting more pleasure from food than you are! The same thing happens when you learn to deeply appreciate certain kinds of art, and then notice that the folks who enjoy "low" art are visibly having more fun.

3) People sometimes get unrealistic dreams and endlessly chase them, like trying to "make it big" in writing or sports, because they randomly got rewarded for it at an early age. I wrote a post about that.

I'm not offering any easy answers here. But it seems like too many people get locked in loops where they spend more and more effort to get less and less happiness. The most obvious examples are drug addiction and video gaming, but also "one-itis" in dating, overeating, being a connoisseur of anything, striving for popular success, all these things follow the same pattern. You're just chasing after some Skinner-box thing that you think you "love", but it doesn't love you back.

Sooo... if you like eating, give yourself a break every once in a while? If you like comfort, maybe get a cold shower sometimes? Might be a good idea to make yourself the kind of person that can get happiness cheaply.

Sorry if this post is not up to LW standards, I typed it really quickly as it came to my mind.