Analysis of superbombs as a global threat

[Draft containing potential infohazards. Do not share without permission of the author.]

Consider ‘superbombs’, that are more destructive than all existing human weapons added together.

Since even stars are sometimes destroyed, we have a proof of concept, and thus every reason to think such weapons are physically possible. Given this, there seems a high chance that they will be developed by us in the course of time if our civilization survives to technological maturity. The question is just how long this will take.

Given that explosive power will eventually vastly exceed all existing human weapons, but the most powerful weapons are currently merely city-scale, one is led to wonder how quickly this transition might take place.

We shouldn’t expect a high degree of precision here, but let us try to distinguish between:

Fast: minutes, hours or days
Moderate: months or years
Slow: decades to centuries

We can roughly think of the situation as a function of two variables: “optimization” power—or quality-weighted effort being applied to the problem—and “recalcitrance”—essentially the effort needed to make progress. We have:

$$Rate\ of\ change\ in\ explosive\ power\ = \frac{optimization\ power}{recalcitrance}$$

First let’s consider recalcitrance. We must be quite uncertain here, since recalcitrance could depend hugely on the final design of such weapons. We can imagine progress becoming more difficult over time, as low-hanging research fruit are taken. But we can also imagine small amounts of effort producing sudden advances, for instance if a person comes up with the single insight needed to arrange available materials into an effective superbomb. Or maybe the different parts of such a bomb would be developed separately, so that building the wick of the bomb is the last bit, and that part is really easy because it’s just a wick, so after a period of apparent stagnation, progress skyrockets.

Our natural tendency to understand explosives at the human military scale may also lead us to overestimate recalcitrance. Having discovered weapons that can destroy a city, we hardly notice

the difference between the first nuclear test's 22kT power, and Tsar Bomba's 50 MT— two thousand times larger. So

Even if bomb design recalcitrance is relatively high, overall recalcitrance might be low, if other components of bomb power have low recalcitrance. Two important aspects of bomb power that might be improved are bomb scale, and bomb fuel.

If we design advanced bombs, but they are initially very small, we might be able to rapidly scale them up, especially if there is huge investment inspired by their promise. We call this scenario a 'resources overhang'.

Similarly, if we design advanced bombs and they are able to make use of different fuels that we have already put centuries of research into perfecting, we may be able to immediately upgrade these new weapons to take advantage of the most highly optimized fuels. We call this a 'fuel overhang'.

In sum, it is difficult to say what the trajectory of recalcitrance will look like, but several different factors could each make it low.

Now let us consider optimization power. Such bombs would be unprecedentedly powerful, and so there are very large incentives to invest. The first team with superbombs will presumably control the world. Arms race dynamics should only increase this further. So we can expect optimization power to climb steeply.

We can also expect a strong feedback effect: perhaps even before superbombs, parties with sufficiently large bombs can begin to use them to overpower others and thus accrue resources, which they can then redirect to explosives research, accelerating progress toward superbombs.

So, recalcitrance could easily be low, and optimization power seems likely to be high and increasing. Thus while it seems possible for superbombs to be developed slowly, most likely the path to superbombs will be moderate to fast and accelerating. In sum, it seems probable that superbomb development unfolds over minutes to years, leaving humanity entirely unprepared and vulnerable to annihilation.