Design, Alignment, and the Structure of Values

The development of AI raises urgent questions about which and whose values it should be aligned with. This has come to be known as the value alignment problem (Russell 2019; Gabriel 2020). Policymakers such as the IEEE or the EU High-Level Expert group are increasingly alert to the issue and call for technology in general, including AI-based technology, to be aligned with or designed for ethical values (EU 2020; IEEE 2019). AI may make the need for an answer more urgent, but the need for technology design to be aligned with human values (dubbed 'design turn' in ethics by van den Hoven et al. (2017)) is much older and rooted in value sensitive design and cognate approaches (Friedman and Hendry 2019). Solving the value alignment problem may be a pre-condition for safe and legitimate development of more advanced AI-based technologies (Aguirre et al. 2020).

Apart from the technical question of how to embed values in technological artefacts by way of design, the normative question of what values – and whose – to align with has attracted some much needed attention. For instance, the value concept in value sensitive design has been criticised as inappropriately descriptive (Manders-Huits 2011), problematically universal (Borning and Muller 2012), and researchers have urged for the need to justify the targets of value alignment with normative theory (Jacobs and Huldtgren 2018; Albrechtslund 2007, p. 67). In the light of disagreement about normative theory, others argued in favour of procedural democratic approaches to determine fairly the values targeted for alignment (Gabriel 2020; Taebi et al. 2014).

These are important contributions because, in different ways, they highlight the normativity of values. Values are normative in the sense that they give us reasons to believe, desire, feel, or act rather than merely (de facto) inciting or motivating us. However, these approaches are marred by normative disagreement, and they lack a metaethical substantiation.

This paper seeks to contribute to the solution of the value alignment problem by introducing an metaethical framework that will help researchers to assess the structure of value. Rather than defending a substantive metaethical view about what values are or a particular normative theory to identify values (e.g. deontology), the paper shows that any solution to the value alignment problem necessitates a normative explanation of the sort 'x is of value because y.' It then shows how to account for such claims by way of different metaethical structures of value. I distinguish *valuing* as a subjective or intersubjective responses from *value*, understood as a normative fact. Though valuing may serve as evidence for value, it must be grounded by a normative claim that associates descriptive properties (e.g. 'is safe') with evaluative or normative properties (e.g. 'is good'). The rest of the paper spells out different possible value structures instantiated by such grounding claims. Some grounding claims are normative laws that explain the normative force of valuing. There may be more fundamental metaethical laws, also expressible as grounding claims, which explain normative laws. An important property of these laws is their stability, which determines how small changes in the grounds affect the grounded values. This, I argue, is of crucial importance the design of technologies aligned with our values.